OULU 2007
C 287

ACTA UNIVERSITATIS OULUENSIS

Ulla Elsilä

KNOWLEDGE DISCOVERY METHOD FOR DERIVING CONDITIONAL PROBABILITIES FROM LARGE DATASETS

FACULTY OF TECHNOLOGY,
DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING,
UNIVERSITY OF OULU

ACTA UNIVERSITATIS OULUENSIS
C Technica 287

ULLA ELSILÄ

KNOWLEDGE DISCOVERY METHOD FOR DERIVING CONDITIONAL PROBABILITIES FROM LARGE DATASETS

Academic dissertation to be presented, with the assent of the Faculty of Technology of the University of Oulu, for public defence in Auditorium TS101, Linnanmaa, on December 14th, 2007, at 12 noon

OULUN YLIOPISTO, OULU 2007

Acta Univ. Oul. C 287, 2007

Supervised by
Professor Juha Röning

Reviewed by
Doctor Bogdan Filipič
Doctor Jukka Kömi

ISBN 978-951-42-8668-1 (Paperback)
ISBN 978-951-42-8669-8 (PDF)
http://herkules.oulu.fi/isbn9789514286698/
ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)
http://herkules.oulu.fi/issn03553213/

Cover design
Raimo Ahonen

OULU UNIVERSITY PRESS
OULU 2007

Elsilä, Ulla, Knowledge discovery method for deriving conditional probabilities from large datasets

Faculty of Technology, University of Oulu, P.O.Box 4000, FI-90014 University of Oulu, Finland,
Department of Electrical and Information Engineering, University of Oulu, P.O.Box 4500, FI-90014 University of Oulu, Finland

Acta Univ. Oul. C 287, 2007
Oulu, Finland

Abstract

In today's world, enormous amounts of data are being collected everyday. Thus, the problems of storing, handling, and utilizing the data are faced constantly. As the human mind itself can no longer interpret the vast datasets, methods for extracting useful and novel information from the data are needed and developed. These methods are collectively called knowledge discovery methods.

In this thesis, a novel combination of feature selection and data modeling methods is presented in order to help with this task. This combination includes the methods of basic statistical analysis, linear correlation, self-organizing map, parallel coordinates, and k-means clustering. The presented method can be used, first, to select the most relevant features from even hundreds of them and, then, to model the complex inter-correlations within the selected ones. The capability to handle hundreds of features opens up the possibility to study more extensive processes instead of just looking at smaller parts of them. The results of k-nearest-neighbors study show that the presented feature selection procedure is valid and appropriate.

A second advantage of the presented method is the possibility to use thousands of samples. Whereas the current rules of selecting appropriate limits for utilizing the methods are theoretically proved only for small sample sizes, especially in the case of linear correlation, this thesis gives the guidelines for feature selection with thousands of samples. A third positive aspect is the nature of the results: given that the outcome of the method is a set of conditional probabilities, the derived model is highly unrestrictive and rather easy to interpret.

In order to test the presented method in practice, it was applied to study two different cases of steel manufacturing with hot strip rolling. In the first case, the conditional probabilities for different types of retentions were derived and, in the second case, the rolling conditions for the occurrence of wedge were revealed. The results of both of these studies show that steel manufacturing processes are indeed very complex and highly dependent on the various stages of the manufacturing. This was further confirmed by the fact that with studies of k-nearest-neighbors and C4.5, it was impossible to derive useful models concerning the datasets as a whole. It is believed that the reason for this lies in the nature of these two methods, meaning that they are unable to grasp such manifold inter-correlations in the data. On the contrary, the presented method of conditional probabilities allowed new knowledge to be gained of the studied processes, which will help to better understand these processes and to enhance them.

Keywords: continuous casting, data mining, feature selection, hot strip rolling, knowledge discovery process

Acknowledgements

This study was carried out in the Intelligent Systems Group at the Department of Electrical and Information Engineering of the University of Oulu, Finland, from 2001to 2006. Fourteen months of the work took place at Arizona State University in the United States.

From Rautaruukki Oyj (for which the marketing name of Ruukki is used hereafter) I would like to thank Paavo Ruha and Tapio Salonpää in particular and all the others who had the persistence and the interest to listen, to comment, to deliberate, and, most importantly, to explain and to teach.

In addition, I would like to thank Steel Dynamics Inc. for providing the data and, especially, Yury Krotov for giving valuable information about the process under study. Thanks also go to two scholars from Arizona State University: to professor Ampere Tseng, who helped with his network of connections and provided the facilities for the research, and to fellow student Ahmed Al-Ghandoor, who helped with the preliminary analysis of the data.

The fourteen months that I spent in Arizona truly meant a lot to me. At this point, I would like to thank Karen Overstake, who became one of my best friends and to whom I owe very much for taking care of me during my stay and, you know, for everything. <LOL>

I am also thanking the Graduate School of Metallurgy and Metals Technology for giving me this opportunity to do my research without having the pressure of the usual short-term projects. The research was funded by the Academy of Finland, the Finnish Cultural Foundation, the Foundation of Technology, Infotech Oulu, the Tauno Tönning Research Foundation, and The Finnish Foundation for Economic and Technologic Sciences, all of which I would also like to thank

Doctors Jukka Kömi and Bogdan Filipič reviewed this thesis and I would like to acknowledge the importance of their work. It has been crucial for me.

For my supervisor Juha: Our work together started in May 1998 and without these years of working with you I do not think that I would have become this more independent and self-confident person that I am today. You have guided me well, even though sometimes I believe I have been the pain in the neck for you.
:-)

My friends Satu Tamminen and Susanna Pirttikangas: What a team we have been! It is so rare that your work colleagues end up being your very good friends. I know that we will stay that way. When I think of my parents and my parents-in-law, I feel that the amount of support that you have given me is unmeasurable. It

is very hard to express all of my feelings for you except by saying that I love you all.

As Riitta's older sister, I have always felt that I should protect her and take care of her, but I have this funny feeling that, at many times, the roles have been vice versa. Well, I do not think it is a bad thing, Riitta. Bisous!

And my daughter Tua, you are my princess. I really want to see if you will become the angel that you say you are going to be when you grow up.

And most importantly for my husband Kalle: You definitely are my better half! I must have done something extremely right to have you by my side. I Love You!

Oulu, November, 2007 Ulla Elsilä

List of symbols and abbreviations

Latin letters

A_1, A_2, A_3	values of class feature
A, B	class labels
c	index for best knot in self-organizing map
C	camber = deviation of the side edge from a straight line
C_j	*j*th value of class feature *C*
d	distance from extreme edge
$f(\ )$	normalization function
$freq(C_j, S)$	amount of samples belonging to class C_j
gain( )	gained information
gain ratio( )	gained proportional information
$h_{ci}(t)$	neighbourhood function between knots *c* and *i* at time *t*
info(*S*)	information that is included in the sample data S
info$_X$(*T*)	information that is included in the sample data T when it is split into subsets with test X
k	number of nearest neighbours
m_c	reference vector for best knot in self-organizing map
m_i	reference vector for knot *i* in self-organizing map
$m_i(t)$	reference vector for knot *i* in self-organizing map at time *t*
max(*x*)	maximum value of feature x
min(*x*)	minimum value of feature x
N	number of samples
n	number of subsets
N/mm^2	unit of measure for tensile strength = Newton / squared millimetre
$p(C_j)$	probability for random sample to belong into class C_j
R^n	*n*-dimensional real vector space
r_i	two-dimensional position of knot *i* in self-organizing map
r_{xy}	estimated correlation of features *x* and *y*
S, T	sample data
s_x	estimated standard deviation of feature x
s_{xy}	estimated covariance of features *x* and *y*
split info( )	possible information that can be achieved by splitting
tc	thickness at centerline

td	thickness at drive side
to	thickness at operator side
w	random feature
X	test for splitting data into subsets
x	input vector for self-organizing map
$x(t)$	input vector for self-organizing map at time t
$\overline{x}$	arithmetic mean of feature x
x_i	i:th value of feature x
$z_{\alpha/2}$	value for standardized normal feature with α level of significance
$\lvert\ \rvert$	amount of samples
$\lVert\ \rVert$	Euclidean distance

Greek letters

α	level of significance
μ_w	mean of w
ρ_{xy}	correlation of features x and y
$\hat{\rho}_{xy}$	estimated correlation of features x and y
σ_w^2	variance of w

Abbreviations

C4.5	decision tree method
CSP	Compact Strip Production
cvc	continuous variable crown
k-NN	method of k-Nearest-Neighbors
OLAP	on-line analytical processing
SOM	self-organizing map

Contents

Abstract
Acknowledgements **5**
List of symbols and abbreviations **7**
Contents **9**
1 Introduction **11**
1.1 Background 11
1.2 Scope of the thesis 15
1.3 Contribution of the thesis 17
1.4 Outline of the thesis 20
2 Knowledge discovery **21**
2.1 Defining the problem 23
2.2 Data Mining 25
2.3 Utilizing the results 26
2.4 Discussion 26
3 Data mining methods **29**
3.1 Conditional probabilities 29
3.1.1 Basic statistical analysis 30
3.1.2 Linear correlation 32
3.1.3 Self-organizing map 34
3.1.4 Parallel coordinates 36
3.1.5 k-means clustering 36
3.2 k-Nearest-Neighbors 37
3.3 C4.5 42
3.4 Data transformation 45
3.5 Discussion 47
4 Targets of application **51**
4.1 Rautaruukki Oyj (Ruukki) 51
4.1.1 Original datasets 53
4.1.2 Most recent datasets 53
4.2 Steel Dynamics Inc. 55
4.3 Discussion 59
5 Conditional probabilities for retentions at Ruukki **61**
5.1 Original datasets 61
5.1.1 Basic statistical analysis 61
5.1.2 Linear correlation and SOM 63

5.1.3 Parallel coordinates 64
5.1.4 k-means clustering 67
5.2 Most recent datasets 71
5.2.1 Basic statistical analysis 74
5.2.2 Linear correlation and SOM 75
5.2.3 Parallel coordinates 80
5.2.4 k-means clustering 82
5.3 Discussion 85
6 Conditional probabilities for wedge formation at Steel Dynamics Inc. 87
6.1 Discussion 93
7 Results of k-NN and C4.5 95
7.1 k-Nearest-Neighbors 95
7.1.1 Datasets 95
7.1.2 Program runs 96
7.1.3 Results 98
7.2 C4.5 103
7.2.1 Datasets 103
7.2.2 Program runs 106
7.2.3 Results 107
7.3 Discussion 111
8 Conclusions 113
8.1 Discussion 113
8.2 Summary 116
References 119
Appendix 125

1 Introduction

1.1 Background

Nowadays, steel products are expected to be more and more polymorphic and both better and more uniform in quality. This has led to a situation where tighter and more precise process control is needed. Fortunately, as the steel manufacturing processes have become mostly automated, new and more precise measurements have become possible with very high frequencies. Thanks to better measuring conditions, the actual process controlling is being improved. However, while many features related to the process itself and to product quality are already measured, the correlations between these features are still largely unknown and, consequently, are poorly utilized in process enhancement. Thus, the problem of handling large databases has arisen, because although information is easy to gather, it is slow and expensive to analyze.

Traditionally, improvements in, for example, a hot strip rolling process have been pursued with mathematical as well as physical models, which consider the process from the view-point of micro-scale (molecular level) as opposed to macro-scale (object i.e. strip level). For example, the modeling of deformation process (Anan *et al*. 1992, Panigrahi 2001) has been the center of attention. The advantages that can be gained by using mathematical and physical models have already been exploited to their near maximum as these very strict models do not take into account the imprecision of the complex processes. Another downside of these models is that the known formulas only include a fractional part of the features that have an effect on the production. Thus, what are needed are methods that can help to discover previously hidden but useful knowledge in large databases.

Methods that can help to study large datasets are called knowledge discovery methods. In scientific literature (Gyenesei 2004, Mannila 1997, Pyle 1999), the descriptions of the process of knowledge discovery are generally quite alike. For example, although Mannila (Mannila 1997) and Pyle (Pyle 1999) differ in the importance they attribute to the different parts of the process, several similarities can be found in their approaches. However, an aspect for which the descriptions vary frequently is the level of automation. Mannila (Mannila 1997) defines the process as an interactive and not fully automated system and Keim (Keim 1995) states that the process of finding information cannot be fully automated since it

involves human intelligence and creativity that are still unequalled by computers. Matheus *et al.* (Matheus *et al.* 1993), 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 far from reach.

The knowledge discovery process of this thesis is presented in Fig. 1. At the beginning, different methods of knowledge discovery and the chosen database 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 prepared and evaluated. These two parts are likely to be done iteratively, because the result from the evaluation probably gives more information about the data and, hence, the preparation of the data can be done more purposefully. The last stage consists in the presentation and the possible application of the results.

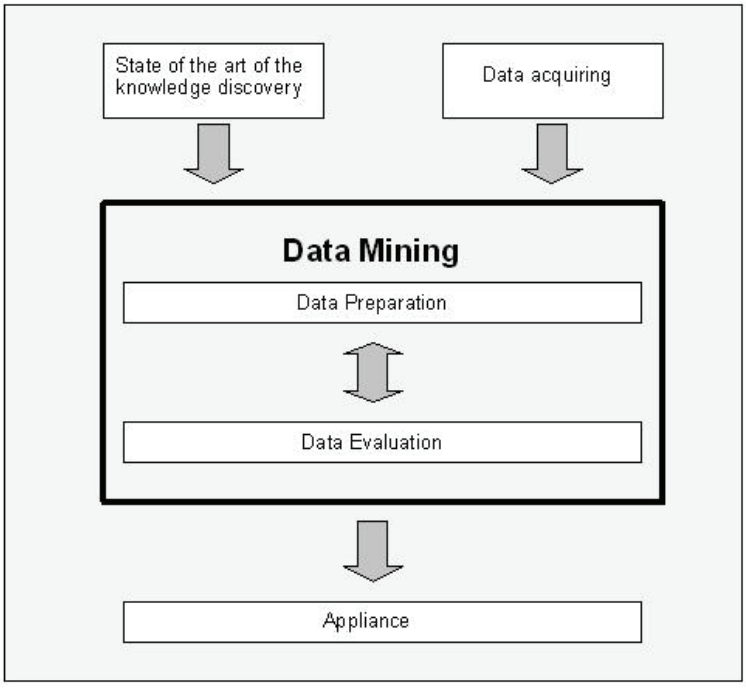

Fig. 1. The main stages of the knowledge discovery process used in this thesis.

As targets of application for this thesis, two different cases were chosen. In the first one, the case of Ruukki, the goal was to discover new knowledge that would enable to anticipate, first, the occurrence of retentions and, second, their absence. It is important to study both of these situations of hot strip rolling, because with the knowledge of circumstances in which retentions appear it is possible to find out what are the causes producing them, and with the knowledge of circumstances in which retentions do not appear it is possible to find out how to produce saleable products. Furthermore, in order to avoid the unwanted outcome one should first know how it is possible to produce one. Herein, the word retention is used to refer to a situation during hot strip rolling in which the product of rolling (i.e. coil) is not automatically accepted for delivery, but has to be retained for re-checking and post-action.

One key quality parameter of a flat steel strip is the uniformity of its thickness across its width, which is called the strip profile (Campos *et al.* 2004). There are many ways that thickness profile can deviate from the uniformity and one of them is called wedge.

In this thesis, the second case was that of the continuous casting mill of Steel Dynamics Inc., where the goal of the study was to discover which features affect the wedge of the strip. At the factory of Steel Dynamics Inc., the wedge is measured right after the finishing mill and, usually, it is within the acceptable limits. Therefore, the actual problem was not to find out what causes the unfavorable products, but to determine how the different parts of the continuous casting process affect the formation of the wedge. This information will broaden the expert knowledge of the process and, if the process should be later changed in some respect, the experts would already have information on how the change could affect the wedge formation.

Today, most knowledge discovery methods are designed to tackle only a selected part of the processes under study. For example, Gyenesei (Gyenesei 2004) searched his data to just find a subset that would then be modeled. Although these types of methods might improve the results obtained of specific parts, the achieved benefits are only a part of the possible improvement that could be gained if the process was studied as a whole. What is more, an improvement in a small part of a continuous process could actually weaken the performance of the following stage. In her thesis, Junno (Junno 1989) states that: “When using a model group that covers the (steel) plant as a whole, local optimizations are avoided, and investments can be more easily arranged according to their economical importance.” Likewise, Bhadeshia (Bhadeshia 1999) states that,

because good engineering is supposed to reach objectives in a cost and time-effective way, any model that deals with only a small part of the required technology is therefore unlikely to be treated with respect. In this thesis, the goal was to find a comprehensive result that would be supported by the whole data.

During the recent years, studying steel manufacturing processes by using various knowledge discovery methods has become common (Bloch *et al.* 1997, Dumortier & Lehert 1999, Femminella *et al.* 1999, Gorni 1997, Portmann *et al.* 1995, Saxén *et al.* 2000). However, the studies have tended to concentrate only on small parts of the processes, which are, for example, galvanization process (Bloch *et al.* 1997), mechanical tensile properties (Dumortier & Lehert 1999), modeling of structure-property relationships (Femminella *et al.* 1999), furnace parameters (Gorni 1997), roll adjustment in advance (Portmann *et al.* 1995), and blast furnace wall temperatures (Saxén *et al.* 2000). In literature, one example was found by Singh *et al*. (Singh *et al.* 1998) where the whole hot strip rolling process was taken into account in predicting the yield strength and the tensile strength of the steel plate.

Two other downsides with the previous studies are that the used databases only include a fractional part of the features that influence the production, and that the rules of selecting right values for models are theoretically proved only for small sample sizes. Previously, the reasons for wedge formation have been studied by, for example, Biggs *et al.* (Biggs *et al.* 2000), Loney *et al.* (Loney *et al.* 2002), Mücke *et al.* (Mücke *et al.* 2002), Shiraishi *et al.* (Shiraishi *et al.* 1991), and Tarnopolskaya *et al.* (Tarnopolskaya *et al.* 2002). However, in all of the above-mentioned studies, only a small number of features were examined. Usually, the selection of a smaller group of features requires expert knowledge about the process, because an inclusion (or exclusion) of only one feature can have an important impact on the results. Sometimes, expert knowledge is not available and, therefore, the method used here was developed so that it can be applied even without a complete understanding of the physical model behind the process. However, the method also allows taking advantage of the knowledge at hand. Also, the aim of this thesis was to discover which features from a more comprehensive dataset influence the wedge of the strip and, thus, the used dataset included features from throughout the rolling process starting right after casting and ending at coiling.

Luckily, a wide range of methods is available today to be used for knowledge discovery. Usually, the study is done with only one method at a time, but this does not necessarily give the best result for problem solving. This is due to the

methods being designed to observe data from only one angle. Furthermore, due to the diverse properties of datasets and the complexity of data mining tasks, no single technique will provide adequate support in all cases (Keim 1995).

1.2 Scope of the thesis

In today's world, it has become somewhat demanding for one person to be an expert in multiple fields. This challenges also people working on knowledge discovering. The main purpose of this thesis is to present a novel method of knowledge discovering, which, in addition to a pure methodological point of view, takes into account the expertise from the owners of the problem. This side of the study is addressed in more detail in Chapter 2.

In earlier studies (Bloch *et al.* 1997, Dumortier & Lehert 1999, Femminella *et al.* 1999, Gorni 1997, Pican *et al.* 1993, Portmann *et al.* 1995, Sbarbaro-Hofer *et al.* 1993, Singh *et al.* 1998), different types of neural networks have been used, but the downside of this method is that it is very complicated to interpret the obtained model. Keim (Keim 1995) suggests that using a combination of multiple techniques helps to produce results that would otherwise be difficult to achieve. Thus, in this thesis, different types of data mining methods were used in order to find correlations between hundreds of measured features. The data mining process of this study is shown in Fig. 2, where the upper rectangular box indicates the methods that were used for feature selection and the lower rectangular box indicates the methods that were used for deriving the conditional probability models.

For feature selection the used data mining methods were basic statistical analysis, linear correlation (Bendat & Piersol 1971), self-organizing map (SOM) (Kohonen 1995), and parallel coordinates (Inselberg 1998, Inselberg 2002). Each of these methods was selected so that it examines different aspects of features and, as a result, this combination makes it possible to handle large datasets with hundreds of features. Because the outcome of the k-means clustering (Vesanto & Alhoniemi 2000) is a set of conditional probabilities, the nature of the model is very unrestrictive as opposed to the nature of, for example, the mathematical model. The outcome is also relatively easy to interpret. The SOM was also used successfully by Saxén *et al.* (Saxén *et al.* 2000), but, in their study, it was only used to model the dependencies whereas, in this thesis, it was used for both reducing the redundancy and modeling the dependencies. In relation to hot strip rolling, Cser *et al.* (Cser *et al.* 1999) started by using the SOM for studying the

dependencies influencing the quality parameters and for the state-monitoring of the hot strip rolling process. Later on, Cser *et al.* (Cser *et al.* 2001) used the SOM as a feature selector for a neural network application that was used for quality prediction in hot strip rolling. The methods used in this thesis are described in more detail in Chapter 3.

From a theoretical point of view, the two targets of application that were chosen had different tasks: whereas the case of Ruukki was used to develop the method of conditional probabilities, the case of Steel Dynamics was used to verify that the derived method would also work with smaller datasets and that it would not be tied to only one type of data.

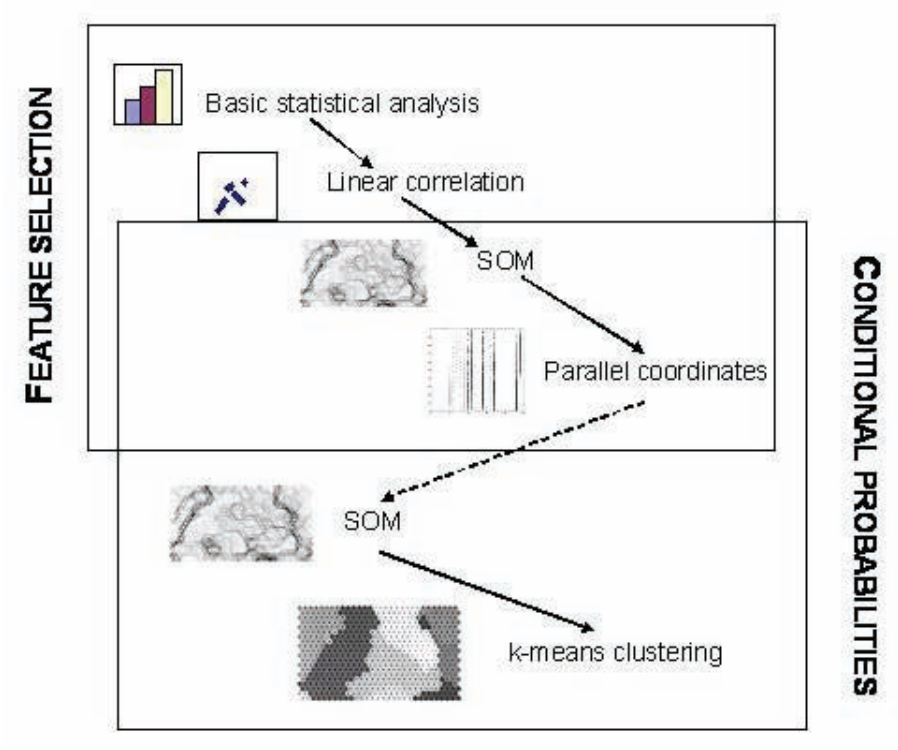

Fig. 2. The data mining process of this thesis.

In order to verify the results of, especially, the feature selection, methods of k-Nearest-Neighbor (k-NN) (Chang *et al.* 2004, Frayman *et al.* 2002, Glick & Hieftje 1991, Laaksonen & Oja 1996, Laine 2003, Shahin & Symons 2001, Tetko *et al.* 2006, Wang & Liao 2002) and C4.5 (Caruana & Freitag 1994, Dash & Liu

1997, Liu & Setiono 1996, Quinlan 1993, Quinlan 1996) were used. Earlier, the method of k-NN has been used to study, for example, genomics (Chang *et al.* 2004), housing (i.e. value of a house) (Frayman *et al.* 2002), classification of alloys (Glick & Hieftje 1991), classification on handwritten digits (Laaksonen & Oja 1996), learning from industrial data (Laine 2003), grading lentils (Shahin & Symons 2001), metal complexation (Tetko *et al.* 2006), and welding defect identification (Wang & Liao 2002). The advantage of both k-NN and C4.5 is that they can utilize the existing class information in the training procedure, but, then again, because they both test individual features at a time, it might be very difficult to identify complex inter-correlations between multiple features. In their article (Liu & Setiono 1996), Liu and Setiono state that, because C4.5 only tests individual features, it is not appropriate for selecting the minimum subset of features. Liu and Setiono tested C4.5 with multiple datasets and the results of their study showed that C4.5 works better with relevant features than with the whole dataset. Similar results are given by Kohavi and John (Kohavi & John 1997) and Caruana and Freitag (Caruana & Freitag 1994). Huang *et al*. (Huang *et al.* 2003) showed that k-NN benefits from feature selection too and Michie *et al*. (Michie *et al.* 1994) stated that the presence of irrelevant features is always a problem for k-NN, regardless of the data size. To add yet another view, Blum and Langley (Blum & Langley 1997) point out that k-NN and C4.5 type of approaches work well in domains where there is little interaction among the relevant features. However, the presence of feature interactions that can lead a relevant feature in isolation to look no more discriminating than an irrelevant feature, can cause significant problems for finding the best solution. It should be also noted that this problem does not disappear with increasing sample size (Blum & Langley 1997).

1.3 Contribution of the thesis

At present, various processes in industry are becoming more automated and complex. This means that process control as well is getting more difficult. As a consequence, knowledge discovery methods are becoming more and more vital, because they can provide a way to comprehend the underlying relationships between the highly numerous process parameters.

The novel ideas and outcomes presented in this thesis are:

- a new way to combine multiple data mining methods to do feature selection,
- the possibility to include more features in original datasets than earlier by using the presented method,
- a novel way to utilize expert knowledge of the problem at hand,
- the use of the special pattern recognition skill of humans at many stages of feature selection,
- guidelines for feature selection with large sample sizes,
- unrestrictive and easily interpretable results, and
- novel information extracted from two different steel manufacturing processes.

The following paragraphs explain the above-listed items in more detail.

This thesis shows a totally new way to combine different data mining methods in order to study large databases and this combination is an effective way to do feature selection from hundreds of features. Such a manifold combination has not been studied earlier and the possibility to use hundreds of features makes it possible to study more comprehensive processes. The use of ensemble models (Frayman *et al.* 2002) is one type of attempt to use different methods to analyze a problem, but instead of using the methods in succession like it was done in this thesis, the ensemble method consists in using different methods to calculate models from the data and, then, combining the achieved models into one or selecting the best model available. Yet, another reason for using feature selection is that the final method that is used to model the data might not be able to handle a high number of features. This might be due to, for example, computational load or the method's inability to handle irrelevant features.

Hence, in this thesis, one main divergence from the earlier studies is that the original datasets are more extensive in respect to the quantity of features under study. Here, the selected processes in question are fast, continuous, mainly automated, and vast numbers of measurements are taken from them with high frequencies. The product varieties are also large and some products are only rarely manufactured. In the case of Ruukki, the target of the application was the whole process of hot strip rolling, starting with slab features measured before heating in furnaces and ending with features measured before coiling. Similarily, in the case of Steel Dynamics Inc., the dataset included features from right after casting all the way until coiling. This selection of larger input space was done

because it is believed that by studying the whole process instead of smaller parts of it, the process will be more extensively enhanced.

A novelty of the presented knowledge discovery method is also the fact that it takes into account expert knowledge from the problem point of view. It was felt that in situations where the process under study is not split into smaller segments, but is, instead, tackled as an entity, it would be quite difficult for the person doing the study to possess all the expertise linked both to the problem and to the method used. Hence, the method presented in this thesis includes the option of using expert knowledge.

There is also another aspect of human skill (besides expert knowledge) that is put into use in the developed method, namely, the ability to perceive patterns. As the selected methods can all visualize data, this inborn skill of humans is utilized at most of the stages of the knowledge discovery process. This skill is hard to mechanize to be as perfect as the human mind.

One of the most important drawbacks of the previous methods is that they provide rules for only small sample sizes, but, with this thesis, guidelines for feature selection with large sample sizes are provided. The presented method offers the possibility to use thousands of samples, instead of just tens or hundreds of them. This should be of interest in the context of most industry related problems, where thousands of measurements have usually already been taken.

Yet another contribution of this thesis is that the results of the presented novel method are not restrictive by nature. The reason for this is that the limits in the resulting conditional probabilities can be overlapping. These results are also easy to interpret when compared to, for example, the results of neural networks that hide the actual decisions inside a so called “black box”.

From the application’s point of view, the purpose of this thesis was to discover new knowledge of the two different cases of hot strip rolling. In the past, the main object to this type of research has been the lack of knowledge of how to utilize the enormous databases resulting from the measurements taken from the rolling process. Now, this problem can be solved with the presented novel method. In the first case under study, the conditional probabilities for different types of retentions were derived and, in the second case, the rolling conditions for the occurrence of wedge were revealed. For both of these cases, the results show that the novel method is fulfilling the task of knowledge discovery process because all the results included some novel and useful information about the processes, in addition to the already known facts by the rolling experts. These

results will broaden the rolling experts' knowledge of the process and, this, in turn, will help them to make better use of the rolling.

1.4 Outline of the thesis

The chapters of this thesis are composed as follows: Chapter 2 presents some theories behind the method of knowledge discovering with special attention given to the steps of problem definition, data mining and result utilization, as well as the role of the experts on the problem. The methods used in this thesis are presented in Chapter 3, which is followed by the presentation of the targets of application in Chapter 4. The Chapters from 5 to 7 present the results derived for three different cases with three different methods. All of the three cases were analyzed with the conditional probabilities method that was developed during the studies. In addition, the first case from Ruukki was analyzed with C4.5 method, whereas the second case from Ruukki was analyzed with the method of k-Nearest-Neighbors. Finally, Chapter 8 gives the conclusions.

2 Knowledge discovery

Today, the method of knowledge discovery is being used in many different environments. In addition to steel manufacturing processes (Bloch *et al.* 1997, Dumortier & Lehert 1999, Femminella *et al.* 1999, Gorni 1997, Portmann *et al.* 1995, Saxén *et al.* 2000) knowledge discovery process has been used, for example, in the analysis of customer relationship management (Rygielski *et al.* 2002), of astronomical data (Fayyad *et al.* 1996), of healthcare data (Matheus *et al.* 1996), of image data (Smyth *et al.* 1996), and of financial markets (Apte & Hong 1996).

What differentiates knowledge discovery process from, for example, on-line analytical processing (OLAP) is that, in knowledge discovery, the information that is searched for is not determined beforehand in contrast to OLAP where the questions are quite strictly fixed in advance (Mannila 1997). This somewhat defines the goal of knowledge discovery: to discover previously unknown information. In scientific literature, the descriptions of the process of knowledge discovery are usually quite alike. Below, two slightly different models are presented. Mannila (Mannila 1997) divides the process into five stages:

1. understanding the domain,
2. preparing the dataset,
3. discovering patterns (data mining),
4. postprocessing of discovered patterns, and
5. putting the results into use.

Mannila states that the process is essentially iterative: the results of the third stage might show the need for a change at the second stage, the fourth stage might steer the user to look for slightly modified models, etc. In his opinion, an effective support for this iteration is a very important factor to be developed in knowledge discovery, as well as the generation of common rules or, in other words, of a theoretical background for the data mining methods.

In order to utilize data in the best possible way, the available information has to be preprocessed. In most cases, the preprocessing itself gives new and useful information about the data. In his book (Pyle 1999), Pyle gives thorough instructions on how to prepare data in order to enable the data mining methods to work in the best possible way. His suggestion for the knowledge discovery process is:

1. exploring the problem space,
2. exploring the solution space,
3. specifying the implementation method, and
4. mining the data

    a) preparing the data,
    b) surveying the data, and
    c) modeling the data.

In this knowledge discovery process, all stages are not equally important. Fig. 3 presents the general durations and importance of the different stages according to Pyle. By looking at this figure it can be seen that the most time-consuming parts are not the most important ones with regards to to the success of the process, but, instead, the least time-consuming part is the most important one. Famili *et al*. (Famili *et al.* 1997) also state that data preparation is a time consuming task.

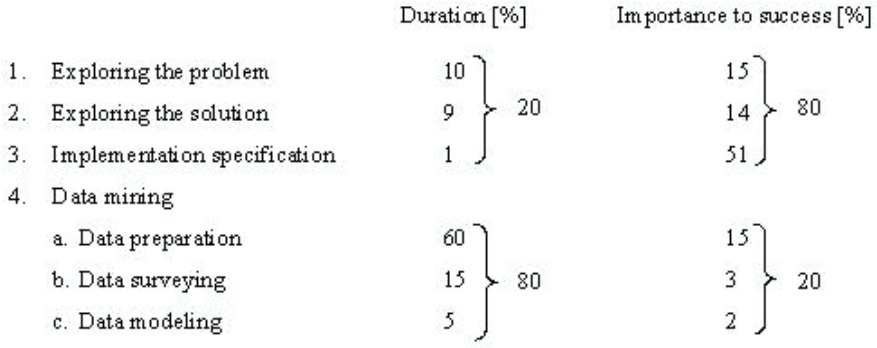

Fig. 3. Stages of a knowledge discovery process showing importance and duration of each stage according to Pyle.

Next, the stages of the process are depicted more precisely. As indicated above, the Pyle's knowledge discovery process starts with the study of the problem space. It is essential to recognize the actual problem in order to achieve useful results. This stage is especially important when a totally new problem is under study. In most cases, the initial description of the problem is not precise enough. It should indeed often be split into smaller parts and, by this means, the solutions obtained could be directed to the correct problem areas.

The second stage corresponds to the study of the solution space meaning that. the desired solution should be defined. The solution could be, for example, a report or an application. The target application, in particular, should be defined

specifically, but the target in question can and should be changed in the course of the process. Yet, it is necessary that the target solution exists from the beginning, because if one does not known what one is looking for, one is unlikely to know if it has been already found. At the third stage, the presented specifying of the implementation method means that it is specifically described how the selected problem solution is going to be used in practice. This is the most important stage of the process in the success point of view.

At the fourth stage, the actual data mining starts and it can be further divided into three parts. At first, the data should be prepared. In Pyle's opinion, the data preparation also prepares the data miner so that he/she can later produce better models in faster pace. The second part is about surveying the prepared data. At this stage, the following questions should be answered: What does the data include? Is it possible to find the answers to the questions with this data? Where are the danger zones? Mainly, this part includes the study of the general structure of the data, in order to evaluate if the data contains enough information concerning the different problem areas. At this stage especially, the iterative nature of the knowledge discovery process becomes very apparent: a discovery of a missing information steers the miner to go back to the data preparation part, where something essential could be discovered with regards to the problem space. Finally, the data is modeled i.e. the optimal answer for a specific problem is sought. For this there is a vast amount of ready methods, such as decision trees and clustering methods. The first achieved results usually give guidance on how the data could be prepared even better and, thus, also at this stage, the iteration is an important way to obtain the best possible solution. In many cases, the process does not end even though a good and useful model has been achieved, because the obtained results usually give a hint of a new problem area and, thus, the process starts again from the beginning.

The knowledge discovery process applied in this thesis is a combination and a modification of the above models by Mannila and Pyle (see also Fig. 1 and Chapter 3.1).

2.1 Defining the problem

Even though the task of defining the problem might seem rather easy at first, it is not. For example, on a factory process line, the workers who control the process might have a totally different idea of the problem as opposed to the factory management although, assumingly, they both hope to produce better products. If

the goal is to obtain better products, should they be better in quality, lower in cost, or something else? Defining the right question might be time-consuming at first, but it will certainly save time and effort at a latter stage of the knowledge discovery process. Pyle gives a good example on the importance of identifying the right problem to solve. In this example, the term “churn” represents the loss of an existing customer to a competitor, which is a problem because the costs of customer acquisition and winning back can be high:

> “In one instance, a major telecommunications company insisted that they had already identified their problem. They were quite certain that the problem was *churn*. They listened patiently to the explanation of the data exploration methodology, and then, deciding it was irrelevant in this case (since they were sure they already understood the problem), requested a model to predict churn. The requested churn model was duly built, and most effective it was too. The company’s previous methods yielded about a 50% accurate prediction model. The new model raised the accuracy of the churn predictions to more than 80%. Based on this result, they developed a major marketing campaign to reduce churn in their customer base. The company spent vast amounts of money targeting at-risk customers with very little impact on churn and a disastrous impact on profitability. (Predicting churn and stopping it are different things entirely. For instance, the amazing discovery was made that unemployed people over 80 years old had a most regrettable tendency to churn. They died, and no incentive program has much impact on death!)
>
> Fortunately, they were persuaded by the apparent success, at least of the predictive model, to continue with the project. After going through the full data exploration process, they ultimately determined that the problem that should have been addressed was improving return from underperforming market segments. When appropriate models were built, the company was able to create highly successful programs to improve the value that their customer base yielded to them, instead of fighting the apparent dragon of churn. The value of finding and solving the appropriate problem was worth literally millions of dollars, and the difference between profit and loss, to this company.” (Pyle 1999)

2.2 Data Mining

In this thesis, the stage of data mining includes two coexistent objects: data preparation and data evaluation (see Fig. 1). It is almost impossible to do data evaluation without data preparation, because, usually and most importantly, the data is not in a form that can be evaluated by the selected method. Also, in most cases, the preparation of the data requires first some kind of evaluation of the data.

A significant part of data preparation consists in the selection of suitable data and, here, the experts from the factory should be involved in order to utilize their knowledge of the problem that is studied. If the dataset is too large, it will take too much time and capacity to compute, but, on the other hand, if the dataset is too small, one may obtain as a result an overfitted model or no results at all. Usually, large datasets are excessive in the sense that two or more features might contain almost the same information. For continuous features, this type of repetition can be located by using correlation analysis (Bendat & Piersol 1971). 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). It has been shown many times (Dash & Liu 1997, Gorni 1997, Hall & Holmes 2003, Setiono & Liu 1997) that the exclusion of the irrelevant, redundant, and noisy features can drastically reduce the running time of a learning algorithm and, what is more, enhance the predictive performance of the resulting model. Dash and Liu (Dash & Liu 1997) also argue that a more general concept is thus achieved.

There are many different techniques to select the features to be studied (Hall 2000, Vafaie & De Jong 1993, Yang & Honavar 1998). Pudil and Novovičová (Pudil & Novovičová 1998) emphasize the idea that there is no unique or optimal way to approach the problem of data subset selection and, in their article (Pudil & Novovičová 1998), they give instructions on how to select a suitable method. Cheeseman and Stutz (Cheeseman & Stutz 1996) warn about the dangers of undocumented and irreversible data preparation. If data is discarded or transformed hastily, one may loose valuable information that could be needed in order to obtain proper results.

After the data preparation stage, begins data evaluation. In this thesis, the methods used for this purpose were SOM, parallel coordinates, and k-means clustering (see Fig. 2). Previously, SOM and parallel coordinates methods have been used separately (Grošelj *et al.* 2004, Inselberg 1998, Kaski 1997, Saxén *et*

al. 2000, Yang & Chou 2003), but not simultaneously as was done here. The experiments done by Vesanto & Alhoniemi (Vesanto & Alhoniemi 2000) indicated that by using the k-means to cluster the SOM instead of directly clustering the data is a computationally effective approach. The clustering results using SOM as an intermediate step were also comparable with the results obtained directly with the data. The order in which SOM and parallel coordinates are used does not have an effect on the results. However, had either one of them been used alone, the comprehension of the resulting rules would have become difficult. By adding the method of parallel coordinates, the results of knowledge discovery are more transparent than when using SOM and k-means clustering without it.

2.3 Utilizing the results

In order to utilize the results properly, it should be first made sure that the results are the answer to the right question or, more precisely, it should be confirmed which question they answer. With this in mind, it is essential that the results are viewed together with the experts in the field of application, because they have the ability to see how the results are linked to the real world. It is extremely rare that the expert facing the problem, is actually doing the knowledge discovery by himself/herself. Thus, the interpretations of both the expert and the person in charge of the knowledge discovery are required in order to properly utilize the achieved results and thereby achieve the final goal of the knowledge discovery process.

Yet, it should be remembered that the possibility of iteration in the knowledge discovery process does not end here. By presenting the results to an expert it can be concluded if appropriate study has been done in the sense that valuable information has been found (Famili *et al.* 1997). At this point, the results might offer a new insight into the problem steering to ask a slightly different question and, thus, give guidance on how to, for example, prepare the data more suitably and efficiently.

2.4 Discussion

This thesis presents a somewhat different view on knowledge discovery process than the earlier models (Mannila 1997, Pyle 1999). As explained at the beginning of Chapter 2, the models by Mannila (Mannila 1997) and Pyle (Pyle 1999)

establish a clear chronological order in which the different parts of the process should be carried out with, at the end, the possibility to iterate the steps if necessary. In this thesis, the underlying difference is that the different steps in the knowledge discovery process are not so clearly separated from each other and, thus, are not necessarily processed independently. Instead, the idea presented here is that some parts of the process are linked in such a way that an individual action is in fact an element common to two parts of the process. For example, an action taken in data preparation (here, SOM and parallel coordinates) is also taken in data evaluation (see Fig. 1, Fig. 2, and Chapter 3.1).

Another divergence from the earlier models is the extension of the problem owner's influence on the knowledge discovery process. It seems that the earlier models utilize the expert knowledge only at the beginning and at the end of the process while, in this thesis, it is suggested that this knowledge should also be put to use during the process. This can definitely enhance the work of the data mining expert as it will broaden his/her knowledge of the problem at hand as well as on the features under study and, more importantly, it will improve the understanding of the results for both the data miner and the problem expert.

3 Data mining methods

This chapter depicts the methods that were used in this thesis. First one to be presented is the developed method of conditional probabilities. In addition, the methods of k-NN and C4.5 are presented, because they were used for verifying the results of the conditional probabilities. Finally, some remarks about transforming the data are made.

3.1 Conditional probabilities

The main steps of the method of conditional probabilities were illustrated in Fig. 1 and Fig. 2. The methods that were used for data preparation or feature selection were expert knowledge, basic statistical analysis, linear correlation, self-organizing map, and parallel coordinates. The data preparation was volotarily comprehensive because, it could be thought that this stage helps to identify interesting subsets from the huge dataset and, thus, makes it much easier for the evaluation method to discover new knowledge within the data. Famili *et al.* (Famili *et al.* 1997) state that complexity may be significantly reduced if irrelevant data are eliminated and only the most relevant features are included in the study. By reducing the dimensionality in this way, the performance of the evaluation method may also improve since the number of training samples needed to achieve a desired error rate decreases with the reduced number of measured features.

Rygielski *et al.* (Rygielski *et al.* 2002) state that data mining software does not eliminate the needs to identify the problem at hand and to understand the data. Hence, it is always important to consult the experts about the data and the results (Famili *et al.* 1997). Earlier, the combination of expert knowledge and statistical analysis has been successfully used, for example, in developing a knowledge-based system for the supervision of a wastewater treatment plant (R-Roda *et al.* 2001). Keim (Keim 1995) states that: “The strategy for dealing with the ever-growing flood of information is cooperation rather than competition between computers and humans. Having people as partners in information interpretation and analysis and allowing them to do what they do best is crucial for making effective use of the available information. In fact, neither computers nor humans alone can solve the kinds of problems that need to be solved in dealing with very large databases.”

In the data evaluation part, the prepared data was, at first, evaluated by using self-organizing map after which the resulting map was clustered with k-means clustering. The result of the clustering was a set of conditional probabilities, which were then analyzed and discussed with the experts.

All of these methods have one thing in common: they all have the ability to visualize the data. Mannila (Mannila 1997) states that visualization is an important technique for mining useful information from large datasets, and it can also be useful in understanding the discovered information. What is more, the visualization allows the user to interact in data exploration (Keim 1995, Keim 2002). Laine used scatter plots and SOM for visualizing and, in his thesis, he states that:

> "Visualization supports the ladder of inference: it supports the analyst to explicate his thinking, both to himself, and to other persons. At the first step of the ladder, the analyst can explicate from the visualized data, which points and variables he has selected for his study. At the second step, addition of meanings, the user can use, for example, histograms to justify his interpretation that a value is high. In the third step the user makes assumptions. He can, for example, use scatter-plots to explicate his strategy of removing outliers, or use the SOM to show how to augment missing values. After presenting his data, his evaluations, and his assumptions, the analyst can proceed to conclusions. Visualization helps to explain and justify the ascension of the ladder, and allows discussion, which justify the ascension in the minds of the audience. The significance of visualization is indirectly proven by scientific literature: most scientific papers contain figures." (Laine 2003)

3.1.1 Basic statistical analysis

It is very natural to do the basic statistical analysis first before applying the other data mining methods. This analysis indeed helps with the data preparation stage at the same time. Hence, the iterative characteristic of the knowledge discovery process is very clearly present here.

At first, the datasets under study in this thesis were selected by using expert knowledge. At second, in conformity with Pyle (Pyle 1999), who states that data preparation starts with a thorough investigation of the qualities of the individual features, each feature was studied statistically in order to determine the types of

features that were included in the dataset. Moreover, Hair *et al.* (Hair *et al.* 1995) point out that 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 of a feature by merely drawing its histogram. In this thesis, 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. In addition, the used datasets themselves included some incomplete sample vectors. The mean and the standard deviation were calculated for each feature, and the following three diagrams were drawn:

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 samples with class value under study.

The second histogram was drawn because the purpose of the study was to find the features that affected on the occurrence of the class value under study. In other words, this was a means to discover the features that had irregularities in their distributions in relation to the classifying feature. Although all this information can be quickly calculated without the need to draw diagrams, visualization was used in order to facilitate the apprehension of the data. The same approach was adopted when drawing the pie charts: the percentages could simply have been calculated or perceived from the first histogram, but the contents of the pie chart are more readily interpreted by the human mind.

As a result of this analysis, the features were categorized in two groups: significant and insignificant features. They were studied one at a time and the rules of the categorizing were as follows:

1. Calculate the percentage of samples with a selected class value in comparison to the total number of samples = class occurrence within the whole dataset.
2. Calculate the percentages of samples within each value or value group (numerical or class) of the feature in question = value occurrence within the whole dataset.
3. Calculate the percentages of samples with a selected class value within each value or value group (numerical or class) of the feature in question = class occurrence within samples with certain feature value.

4. Find the values with a value occurrence (rule 2) of at least 3% and whose percentages calculated according to rule 3 differ at least 2% of the class occurrence calculated according to rule 1.
5. If there is at least one value on both sides of the class occurrence that comply with the fourth rule, the feature in question is categorized as a significant one. Otherwise, the feature is insignificant.

The selection of the percentage limits in rule 4 was based on the results of the first case of Ruukki. This selection is addressed more closely in Chapter 5.1.1. In addition, Chapter 7.1.3 presents verification of the appropriateness of the limits.

The following example furhter illustrates how this categorizing works. A dataset consists of 10000 samples which are classified into two groups: either with label A or B. The label B stands for the unwanted products and the amount of samples classified as B is 1100. Thus, the first rule gives the percentage of 1100/10000 = 11%. Moreover, there is a feature with two values, zero and one, and it corresponds to 7500 samples of value zero from which 1000 samples have class value B and 2500 samples of value one from which 100 samples have class value B. Hence, according to rule 2 we have 7500/10000 = 75% of samples with value zero and, correspondingly 25% of samples with value one. Next, by using rule 3, it is concluded that the percentage of samples with feature value zero and class value B is 1000/7500 = 13.3% and the percentage of samples with feature value one and class value B is 100/2500 = 4%. Now, considering rule 4, both feature values are present with an occurrence of over 3% (75% and 25%) and both class occurrences of these feature values (13.3% and 4%) differ from the class occurrence of the whole dataset (11%) by more than 2%. So finally, the feature is categorized as a significant feature according to rule number 5.

3.1.2 Linear correlation

After the statistical analysis, the datasets were studied by using linear correlation in order to reduce the repetitious information they contained. The necessity of repetition reduction depends on the method that is used to explore the data after its preparation. For example, results acquired with SOM will be erroneous if these excessive features are not excluded. However, correlation can only be calculated for continuous features and, therefore, some of 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 (see an example in Chapter 5.2.2, Fig. 18).

When using correlation values to reduce the number of features, an appropriate limit for exclusion has to be selected. For small sample sizes a test for statistically significant correlation can be found in Bendat & Piersol and it is presented below. First, the estimated correlation for N samples is calculated from

$$r_{xy} = \hat{\rho}_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{xy}}{\sqrt{\left(\sum_{i=1}^{N} x_1^2 - N\bar{x}^2\right)\left(\sum_{i=1}^{N} y_1^2 - N\bar{y}^2\right)}}, \quad (1)$$

where x_i and y_i are the ith samples and $\bar{x}$ and $\bar{y}$ are the arithmetic means of features x and y. Then, the accuracy of the estimate in Eq. 1 can be evaluated with the following test:

Assume zero hypothesis H_0: $\rho_{xy} = 0$, which means that there is a significant correlation, if the hypothesis is rejected. Then, form a random feature

$$w = \tfrac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] \quad (2)$$

which has an approximately normal distribution with a mean and a variance of

$$\mu_w = \tfrac{1}{2}\ln\left[\frac{1+\rho_{xy}}{1-\rho_{xy}}\right] \quad (3)$$

$$\sigma_w^2 = \frac{1}{N-3}. \quad (4)$$

From Eq. 3 and 4, the sampling distribution of w given $\rho_{xy} = 0$ is normal with a mean of $\mu_w = 0$ and a variance of $\sigma_w^2 = 1/(N-3)$. Hence, the acceptance region for the hypothesis of zero correlation is given by

$$-z_{\alpha/2} \le \frac{\sqrt{N-3}}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] < z_{\alpha/2}, \quad (5)$$

where z is the standardized normal feature. Values outside the interval in Eq. 5 would constitute evidence of statistical correlation at the α level of significance. (Bendat & Piersol 1971)

Unfortunately, the above test does not work well when the number of samples is large. For example, if $N = 1000$ and $\pm z_{\alpha/2} = \pm 1.96$ with a level of significance of 5%, then with $r_{xy} = 0.07$ the value $\sqrt{N-3}\, w = 2.21$ falls outside the acceptance region. Hence, a quite minimal correlation of 0.07 is reflected as an existence of a statistically significant correlation. At present, there are no statistical methods to calculate the limit when the number of samples is large. Thus, the limit for correlation has to be chosen separately for each study. It is recommended to use visual inspection of scatter plots while making the decision. In this thesis, the selected limit for exclusion was 0.9, because with inferior values, the information contained in the features was found not to be redundant. Also, taking into consideration the large number of samples used in this thesis, having more than 10% of them uncorrelated is surely an interesting deviation to study. In Chapter 5.2.2, this selection of the exclusion limit will be explained more closely by using a couple of examples.

3.1.3 Self-organizing map

The method of self-organizing map (SOM) was developed by Kohonen (Kohonen 1995), and a very thorough comparison between SOM and other clustering methods has been proposed by Kaski (Kaski 1997). SOM is a very commonly used method in various problems (Cser *et al.* 1999, Cser *et al.* 2001, Cser *et al.* 2001, Goser 1997, Grošelj *et al.* 2004, Saxén *et al.* 2000, Tryba & Goser 1991, Yang & Chou 2003). 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, where each knot includes a various amount of samples that are very similar to each other. This similarity is measured by Euclidean distance. In other words, SOM is a projection of the multidimensional density function into the 2-dimensional space. The clusters are finally formed by classifying the knots into larger groups. In a way, the resulting map is extremely general, considering that beforehand assumptions about the shapes and the number of clusters do not need to be made. The horizontal and vertical axes of the map should not, however, 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. Examples of SOMs and component planes can be found under Chapters 5.1.2 and 5.1.4. In the next paragraph, the calculation of SOM is presented in more detail.

In a self-organizing map, for every knot *i* there is a reference vector $m_i \in R^n$. At the beginning, the reference vectors are initialized. After this, the net is trained at two different stages: first, there is a rough training that defines the global order of the map and, then, the fine-tuning that will provide the map with its final, accurate state. When an input vector is being processed, it is compared to all reference vectors and the knot having the best reference vector will win the input vector. In order to determine the best knot, the Eq. 6 is used. Assume $x \in R^n$ to be an input vector and let the best knot be referred with index *c*:

$$\begin{aligned} \|x - m_c\| &= \min_i \{\|x - m_i\|\} \quad \text{or} \\ c &= \arg\min_i \{\|x - m_i\|\}, \end{aligned} \tag{6}$$

where $\|x - m_i\|$ is the Euclidean distance. During the training, knots that are within a certain distance from each other, activate each other to learn from the same input. Let r_i denote the two-dimensional position of the knot *i* in the map. Learning is described with

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left[x(t) - m_i(t)\right] \tag{7}$$

where $h_{ci}(t) = h(\|r_c - r_i\|, t)$ is the neighborhood function, $\|r_c - r_i\|$ is the radial distance between the knots, and *t* denotes the time. When the radial distance between the knots increases, the value of the neighborhood function h_{ci} approaches zero. This neighborhood function defines the robustness of the surface to be fitted to the inputs and its form can be chosen between the following two: “monotonically decreasing” or “inverse relation to time”. If there are missing values in input vectors, these are not taken into account when the distances between vectors are being calculated. Similarly, when the reference vectors are being modified, only the components that are present are taken into account. It is wotrh noting that, in all occasions, the reference vectors include all the components.

3.1.4 Parallel coordinates

The method of parallel coordinates (Inselberg 1998) is a common data visualization method (Andrienko & Andrienko 2004, Chen & Wang 2001, Chou *et al.* 1999, Fua *et al.* 1999, Hall & Berthold 2000, Siirtola 2003, Weber & Desai 1996). It is a method of transforming the search of relations within multivariate datasets into a 2-dimensional pattern recognition problem. Its major strength is in modeling relations between features. Plainly put, in parallel coordinates, the horizontal axis represents the features and the vertical axis their normalized values. In this way, the distributions of the values of the features are readily compared to each other by looking at just one figure. An example of a parallel coordinates figure is presented in Chapter 5.1.3. The problem with this type of presentation of parallel coordinates, where only one circle or dot presents all the samples with the value in question, is that the density of the value is not shown. This, in a way, leads to a favoritism of continuous components in the selection, because the class components will be more readily excluded from the further studies. This unequal consideration of components should be, at least, taken into account when interpreting the results. On the other hand, if one draws lines to illustrate each of the thousands of samples, the physical size of the figure becomes huge and, in this thesis, the capacity of the computer that was used was not enough to handle this type of figures. In addition, it is very unlikely to find the small, interesting features from it. Keim (Keim 1995) concluded that parallel coordinates is very useful for relatively small datasets with large dimensions. Nowadays, there are multiple ways to enhance the use of parallel coordinates (Andrienko & Andrienko 2004, Chen & Wang 2001, Chou *et al.* 1999, Fua *et al.* 1999, Hall & Berthold 2000, Inselberg & Avidan 1999, Siirtola 2003) and these should be considered in future studies.

3.1.5 k-means clustering

The method of k-means clustering is a model-free prototype method, where the prototypes represent the training dataset by a set of points in a feature space. The prototypes are typically not examples from the training samples. Each prototype has an associated class label and a sample is classified with the closest prototype. Closeness is usually defined by Euclidean distance as was the case in this thesis. In unsupervised k-means clustering, the first thing to do is to select the number of prototypes (i.e. clusters). After that, the training samples are linked to the closest

prototypes and the total variance is minimized in relation to the training samples by iteratively moving the prototypes to the centers of clusters formed by the training samples. This is done until the convergence has been achieved. The problems with k-means clustering are that it usually converges to local optimum and that it is sensitive to outliers (see Chapter 3.4). (Hastie *et al.* 2001)

There are many applications where k-means clustering can be used (McCombie *et al.* 2005, Vesanto & Alhoniemi 2000) and it can also be used for feature selection (Lin *et al.* 2004). In this thesis, at the final stage of deriving the rules, the resulting SOMs were clustered by using a k-means clustering algorithm as Vesanto & Alhoniemi (Vesanto & Alhoniemi 2000) suggested. Clustering means that a dataset is partitioned into groups of samples, where samples within one group are similar to each other. In k-means clustering, the first step is to determine the number of clusters, after which the dataset is partitioned. In this thesis, the number of clusters was unknown beforehand and, thus, the k-means clustering was repeated so that the number of clusters increased from 2 to 25. The actual calculating was done with Matlab, where the k-means function was run multiple times for each number of clusters, and the best of these was selected based on sum of squared errors. Next, the Davies-Bouldin index (Davies & Bouldin 1979) was calculated for each clustering, after which the best possible result of clustering, i.e. the optimal number of clusters, was selected according to an index resulting of a function that includes within-cluster distances and distances between the clusters. The best clustering has the smallest Davies-Bouldin index. Finally, the whole process was repeated ten times and the clustering with the smallest Davies-Bouldin index was selected. The purpose of these repetitions was to maximize the odds to find the global optimum. Now, when the SOMs were clearly divided into different clusters, the conditional probabilities could be derived for each of them. An example of clustering is presented in Chapter 5.1.4.

3.2 k-Nearest-Neighbors

The reason for selecting k-Nearest-Neighbors for verification method was that it is well known and widely used (Chang *et al.* 2004, Frayman *et al.* 2002, Glick & Hieftje 1991, Huang *et al.* 2003, Laaksonen & Oja 1996, Laine 2003, Mitchell 1997, Shahin & Symons 2001, Tetko *et al.* 2006, Wang & Liao 2002). With the method of k-NN, the features are selected by using the classifier that later on uses these selected features in predicting the class labels of unseen samples. The

classification of an unseen sample is done, first, by finding the k nearest training samples and, then, by classifying the sample according to the majority of the classes. Usually, the used distance measure is Euclidean distance and, hence, the features should be scaled to be commensurable.

The characteristics of this kind of method are that the selected subset of features is not suitable for different classifiers and that the computational load is very costly, but, on the other hand, the accuracy level is very high. The method of selecting the features plays an important role in the search of the best possible prediction. Because of the high computational load, the feature base is not usually searched exhaustively in order to find the best combination of features, but, instead, some restrictions are made. For example, if there were 70 different features, the number of different combinations without repetition would be approximately 10^{21}. Hence in this thesis, to overcome the extensiveness of calculations, the following feature selection methodology was used (see Fig. 4 and Fig. 5): First, one feature was selected so that it gave the best prediction on training (i.e. development) dataset, and second, all combinations of two features including the previously selected feature were tested and the combination giving the best prediction on training dataset was selected. It is worth noting that all possible combinations of two features were not studied, but only the ones that included the previously selected feature. The same analogy applies through the whole selection process meaning that, for any given number of features to be included in the combination, all possible combinations were not studied, but the calculations were only done with the combinations that included all the previously selected features plus one. It was understood that this type of selection methodology left a large amount of possible combinations unstudied and, thus, a limited loop back for extracting single features was included. With this, it was made sure that one feature could not control the whole selection procedure. After the third feature was selected, one feature from the selected combination was extracted if this yielded a better prediction. The possibly extracted feature was not removed from the feature base, but was simply returned to the set of unstudied features. Because of the possibility of an eternal loop, this method of extraction was limited to three times for any given number of features to be included in the combination. All in all, the stages of selection and extraction were repeated until all the features were included in the final combination and all the results of predictions for both the training and the evaluation datasets with the tested combinations were saved for further analysis.

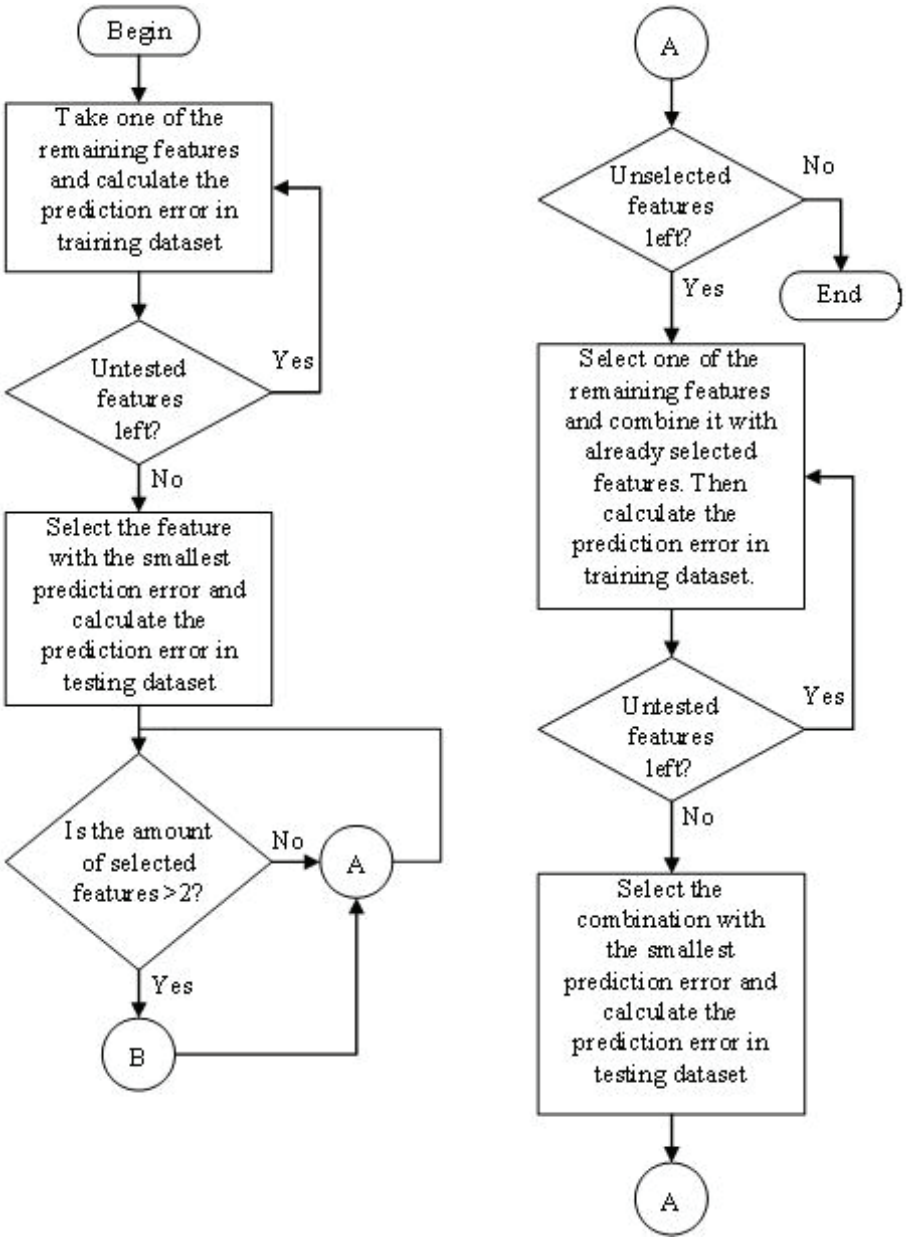

Fig. 4. The flowchart of the used k-Nearest-Neighbor program. (cont.)

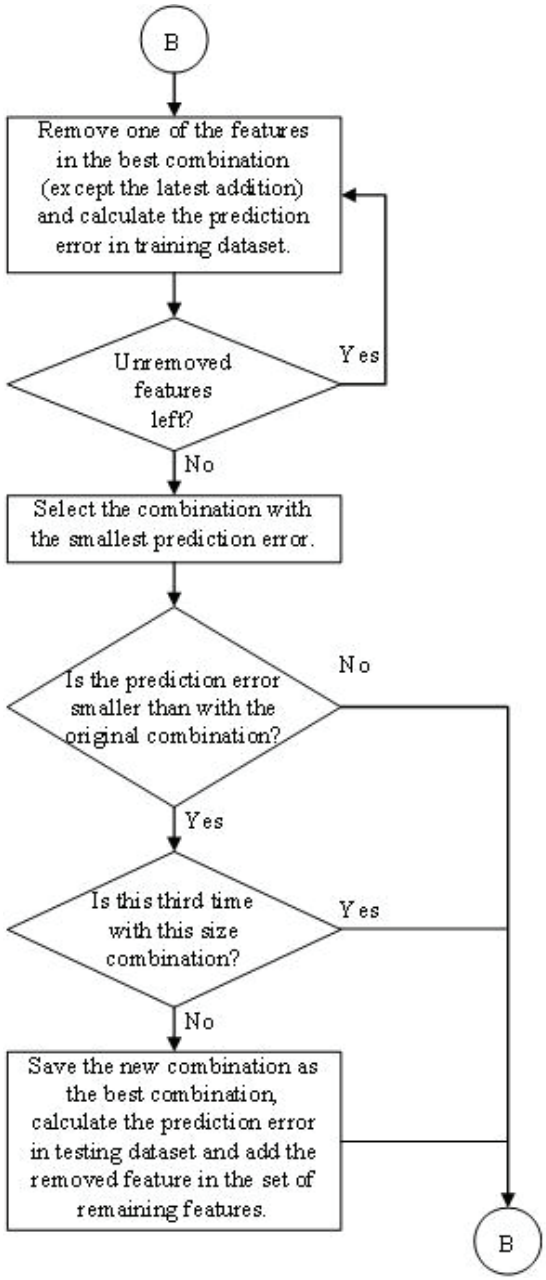

Fig. 5. (cont.) The flowchart of the used k-Nearest-Neighbor program.

In this thesis, three different types of measures for selecting the best prediction with training dataset were tested:

1. proportional error of prediction on the whole dataset,
2. proportional error of prediction on the samples with retentions, and
3. arithmetical average of two values: the proportional error of prediction on the samples with retentions and the proportional error of prediction on the samples without retentions.

The feature to be added or extracted was then selected so that it gave the smallest value for the measure over all given alternatives of feature combinations. The analogy behind using these different measures was that the feature selection procedure might perform in very different ways with different measures, because the datasets were not equally balanced with samples of both classes i.e. samples with and without retentions. For example, let us assume that there is a dataset with two classes, namely, A_1 and A_2, where 90% of the samples belong to class A_1 and 10% to class A_2 in both the trainig and the evaluation dataset. Now, if the best prediction is measured by using the proportional error of prediction on the whole dataset, it is possible that a very good prediction, e.g. 90% correct, is gained just by classifying all the samples to belong into class A_1. Another motive for using other measures than the proportional error of prediction on the whole dataset was that good prediction results for samples with retention (i.e. the smaller class) were of special interest and, hence, two other measures with focus on this class were created.

Because the k-NN classifier is the evaluation function of the classification, the method is called a wrapper method (Dash & Liu 1997). The method is also referred to as an unsupervised method, meaning that it does not make any assumptions on the data and, is, thus, suitable for all situations. On the other hand, any given part of the decision boundary is only dependent on a few (=k) samples and their distribution, which renders the boundary wavy and unstable. This results in great variance and small bias. By increasing the number of neighbors the variance can be decreased, but, at the same time, the bias will increase. The best number of neighbors, i.e. the best value of k, is found by testing different values for k and selecting the value that minimizes the classification error with the evaluation dataset. It should also be noted that all of the k nearest neighbors have an equal impact on the prediction, which can lead to an unstable class distribution. Hastie *et al*. (Hastie *et al.* 2001) conclude that it seems that with the k-NN method, the error on the training dataset should be approximately an

increasing function of k, and is always zero with k=1. In this thesis, values 3, 5, and 10 were used for k.

3.3 C4.5

The most important advantage to be gained with the recursive binary tree is the easy interpretability of the results (Hastie *et al.* 2001). The goal of the method is to distribute the data in box-like areas in such a way that each area includes samples with only one class value and that the samples of each area are in the best possible way distinguished from the samples of the other areas. Usually, a perfect situation such as this is not achievable and, thus, some restrictions for the search are used in order to reduce the computational load. What is more, the resulting tree can be later pruned to better meet the problem.

C4.5 is a group of programs that use inductive generalizations to build a classification model (Quinlan 1993, Quinlan 1996). In order to conduct classification in an inductive way, some criteria have to be met:

- Feature values: Each sample has to have a constant number of feature values. A feature may include discrete or numerous values, but the features used to describe the samples cannot vary from one sample to another.
- Pre-defined classes: The classes, into which the samples are divided, have to be predefined.
- Discrete classes: The classes have to be strictly defined, i.e. a sample either belongs or not into a certain class and all the samples have to be classified. In addition, there should be much more samples than classes.
- Sufficient dataset: Sometimes, a simple model can be identified with only a handful of samples, but, in most cases, a detailed classification model requires hundreds or even thousands of samples.
- "Logical" classification models: The programs only produce classifiers that can be depicted with decision trees or production rules. These limit the description of classes into logical expressions whose conditions are clauses of certain feature values.

In this thesis, every dataset fulfilled these criteria: there were no missing values, the samples had clear class labels and the classes were predefined, thousands of samples were available, and the models could be described with logical expressions.

The whole system derives its name from the program C4.5 that generates a classifier in the form of a decision tree. In the tree, a leaf represents a class and a decision node defines a test for a given feature value. There is one branch or subtree for each possible outcome of the test. At the beginning, the program constructs an initial decision tree using training samples. Usually, the tree is very complex and overfitted and it should, thus, be simplified. There are two ways to do this: either by testing the possible gain of subtrees before they are build or by recursively simplifying an overfitted tree. The former allows time not to be wasted on building a structure that will not be later used in the simplified tree, but the problem is that it is very difficult to define the threshold, starting of which the subset should be further divided. The latter is the method used by C4.5 and although it consumes more time, it is, at the same time, more reliable. The recursive simplification can be done by pruning, which means that the parts of the tree that do not improve the predicted classification accuracy concerning the test samples are removed. All in all, with the simplification, or pruning, a less complex and more easily understandable tree is obtained. (Quinlan 1993) In addition, Michie *et al*. (Michie *et al.* 1994) stated that pruning enhances the results.

Besides pruning, the following procedures were used in this study: grouping of feature values, soft thresholds and windowing. When the algorithm encounters a discrete feature, the default rule is that it creates a different branch for each feature value. In the case of multiple feature values, this might lead to two problems: subsets within branches might be too small for finding useful rules and the feature values might be treated unequally, if their subset sizes differ significantly. The number of the branches can be cut down by using the grouping of the feature values.

In the case of the continuous features, if a value is close to the threshold of the test, a small (insignificant) change in the value might produce a radical change in classification. This can be avoided by using the procedure of soft thresholds.

The third procedure, i.e. windowing, is an indirect way to construct a decision tree from large datasets. For the initial window, C4.5 chooses training samples so that the class distribution is as uniform as possible. The tree constructed from these samples is used to classify the training samples outside the window and at least half of the samples that were classified incorrectly will be selected into the next window. The cycle continues until all the samples outside the window are correctly classified but the cycle can also be interrupted, if it seems that the

classification accuracy is not improving. Here, the improvement is simply measured based on the quantity of incorrectly classified samples.

Yet another procedure that can be selected when making C4.5 program runs is the selection between gain criterion and gain ratio criterion, which are ways of measuring the quality of tests that are located in the decision nodes of a tree. The gain criterion is defined as follows:

The probability for a random sample to belong to a class C_j is

$$p(C_j) = \frac{freq(C_j, S)}{|S|}, \tag{8}$$

where S is the sample data, $|S|$ is the quantity of samples, and $freq(C_j,S)$ is the number of samples that belong to a class C_j. Thus, the information that is included in the sample data with k classes can be measured with:

$$\text{info}(S) = -\sum_{j=1}^{k} p(C_j) \log_2 \left(p(C_j) \right) \text{ bits.} \tag{9}$$

Now, info(T) measures the average information needed to identify a class for a sample in data T. When T is split into n subsets with test X:

$$\text{info}_X(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \text{info}(T_i), \tag{10}$$

the gained information is:

$$\text{gain}(X) = \text{info}(T) - \text{info}_X(T). \tag{11}$$

The gain criterion chooses the test that maximizes the gained information and it tends to prefer tests with multiple outcomes. This can be fixed by using a sort of normalization:

$$\text{split info}(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right). \tag{12}$$

This split info represents the possible information that can be achieved by splitting the data T into n subsets, whereas the gain measures the information that relates to the actual classification that comes from the same split. Now,

$$\text{gain ratio}(X) = \frac{\text{gain}(X)}{\text{spilt info}(X)} \tag{13}$$

indicates the proportional information that is achieved from the split. This information is useful for the classification. The gain ratio criterion chooses a test that maximizes the gain ratio so that the gain is high – at least as high as the average gain from all the tests that have been studied. This limitation insures that the split is not trivial (when the split info is small and the ratio is unstable). Quinlan thinks that the gain ratio criterion is robust and it generally produces better choice of tests than the gain criterion. It even appears to be advantageous when all the tests are binary and differ by the numbers of samples in the two outcomes. (Quinlan 1993) In addition, Caruana and Freitag (Caruana & Freitag 1994) concluded that the performance of the gain ratio criterion was better than the performance of the gain criterion when no feature selection was made.

3.4 Data transformation

In this thesis, both the SOM and k-NN methods use the Euclidean distance measure. In the case of SOM, the used software did not allow to choose any other measure than the Euclidean distance and, in the case of k-NN, this measure was the obvious first choice because of its universality. Of course, testing other distance measures with k-NN could have been possible, but, within the scope of this thesis, this was not done.

Given that the used measure is the Euclidean measure the features have to be scaled (or normalized) so that they have equal intervals. In practice, this means that the weights of the features will be equalized and, thus, the underlying assumption is that all the features are similarly important (Lumijärvi 2003). In other words, all the features have the same impact on calculating the distances between different samples. This is not always the case, but, within the scope of this thesis, it is assumed to be so, because the idea behind the developed method is that one does not need beforehand knowledge to assume otherwise. Another view that had to be considered was the selection between scaling and normalization by using the average and the variance. If the feature is not by its nature N-distributed, then the latter way would change its nature. Instead, when the difference between maximum and minimum values of the feature is used, the distribution of the feature remains also after the scaling. In this way, the correlations stay the same too. Hence, in this thesis, the features were actually scaled to be commensurable and not normalized in the strict sense of the word. However, for the sake of the subtle difference between the terms, the term of

normalization is used, because it seems to be more widely employed in technical environments.

In the context of SOM, the features are referred to as components. Within this thesis, the difference between features and components is that features include the true sample values and components include the normalized sample values. Normally, the used interval for normalization is either [0,1] or [-1,1] and, when the Euclidean distance measure is used, both of the intervals should give similar results. In this thesis, the former interval was selected, because the datasets included also class features, which where thus more feasibly transformed into components (see below). It is easy to normalize continuous features and, for this purpose, the Eq. 14 was used:

$$f(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \tag{14}$$

where x_i is the ith sample of feature x, min(x) and max(x) are the minimum and the maximum values of feature x. With this type of normalization, there is one problem: the outliers in data tend to dominate the normalization and, thus, the feature values are usually divided by standard deviation (Lumijärvi 2003). In order to avoid this, Wilson and Martinez (Wilson & Martinez 1997) used the division by four standard deviations to scale each value. This selection of the divisor was derived from the fact that in normal distribution, 95% of the values fall within two standard deviations of the mean, thus leaving 2.5% of the values for each tail of distribution. In this thesis, large values of one of the features (*duration of slab heating*) were all given a normalized value of 1. As the feature in question did not have an ideal normal distribution, the maximum value was selected so that approximately 2.5% of the values were outside this limit. Within each subset, the maximum to be used for that feature was three times its deviation added to its average. Consequently, at least 97.4% of the data was treated normally, meanig that the corresponding values were less than the selected maximum.

In the case of class features, the task of normalizing becomes slightly more time-consuming. In this thesis, for every class feature, new components were created so that every component represented only one class. For example, if a feature included classes A_1, A_2, and A_3, the first new component would have value one, if the sample belonged to a class A_1, and value zero otherwise. Correspondingly, the second new component would have value one, if the sample belonged to a class A_2, and zero otherwise. The third new component would have

value one, if the sample belonged to a class A_3, and value zero otherwise. Thus, a feature with three class values would yield three components, four class values would yield four components, and so forth. In the case of only two class values, only the first component was used, because the other one would have been the exact complement of the first. Moreover, in cases with more than two class values, one of the components could have been left unused, but then the information of that component would have been hidden behind a methodically complicated union of other components. This would not have been desirable.

3.5 Discussion

This thesis describes a novel method of feature selection and data modeling. Its usability lies in several aspects:

1. the data can be processed with or without expert knowledge of the problem (reader should be informed that the use of such knowledge is highly recommended),
2. the dataset under study is allowed to be voluminous with regards to the number of both features and samples,
3. each of the used methods has a rather strong capability to visualize data, and
4. the results in form of conditional probabilities are very unrestrictive and relatively easy to interpret.

Another strength of the present method is that by using multiple feature selection and data mining methods in succession, it is possible to examine different aspects of the features and the problem at hand at the same time. This is essential since these two are usually linked to each other.

The choice to use basic statistical analysis as the first method was rather obvious. It is a very effective way to become familiar with the data and it can even offer novel information to the process experts involved about what is happening in the process. In order to use this method for feature selection, some guidelines and limits have to be given. In this thesis, the basic statistical analysis was used to categorize the features into two groups: significant and insignificant features. The process of categorizing is a five-step procedure and it uses some innovative limits for partitioning. The given values for limits worked very well in all cases under study and, thus, they can be used as a starting point for any new problem.

Previously, the problem with using linear correlation as a feature selection method has been that there were no tests for proving insignificant correlation when the sample size was large (i.e. over approx. 100 samples). In this thesis, new guidelines for using linear correlation as a feature selection method are provided. And, as with basic statistical analysis, the given value for a limit is recommended to be used as a starting point.

The human mind has a high ability to recognize patterns when one is observing images or just keeping one's eyes open. But, there is a limitation in the number of dimensions that can be grasped simultaneously. In this thesis, this ability was put into use in feature selection together with the SOM. The underlying idea was quite simple: first, to reduce the dimensionality of the problem by using SOM to visualize it in a 2-dimensional space and, then, to use the human expertise to locate the redundancy in the data. Actually, this is a relatively safe way to reduce the number of features, because, after having excluded some of them, it is possible to use SOM again, compare the resulting maps and verify if there was a change into better or worse.

The parallel coordinates is yet another powerful method for visualization. However, regrettably, all the capabilities of this method could not be utilized in full because of the large amount of samples in datasets. This is definitely a development target for further studies in order to even further enhance the quality of the presented method.

The method of SOM was also used at the final stages of the data mining process to help to produce the conditional probabilities together with the k-means clustering. The usability of combining these two methods was already studied by Vesanto & Alhoniemi (Vesanto & Alhoniemi 2000) and their result was utilized here in a straightforward manner. As the final output of the presented method, this stage produced the conditional probabilities, which were then interpreted together with the experts on the problem. It should be noted that, at the stage of reading the results, expert knowledge can have an important role and, thus, should not be ignored lightly.

The methods of k-NN and C4.5 were presented here in order to verify the results of, especially, the feature selection. As these methods were not the main focus of this thesis, their use was restricted into a couple of common settings. In the future, as the method of k-NN seems to derive better predictions on the dominant class in a dataset, it might be worthwhile to reform the dataset so as to contain equal proportions of different classes. However, then the challenge is to determine a way to obtain a representable set of samples from the dominant class

without loosing any diversity. Another line of further study on the k-NN could be the selection of different distance measures.

4 Targets of application

Two different cases from steel industry were studied. The first case was Ruukki and its hot strip rolling process. The database under study included features starting with slab dimensions measured before heating in furnaces and ending with features measured before coiling. The goal was to discover new knowledge with which the occurrence and absence of retentions could be foreseen.

The second case was Steel Dynamics, Inc. and its continuous casting hot strip rolling process. In this case, the database included features starting with those formed right after casting and ending with those formed during coiling. The goal of the study was to discover which features affect the wedge of the strip.

4.1 Rautaruukki Oyj (Ruukki)

The data was gathered from the hot strip mill of Ruukki in Raahe, Finland. The mill is illustrated in Fig. 6. Before the rolling process, the slabs are fabricated at a steel plant. Slab dimensions, chemical composition of slab as well as the target values of the rolling process etc. are transmitted from the production planning computer to the process control system. Before rolling, the slabs are heated in reheating furnaces. At the time of the collection of the datasets, the mill used one walking beam furnace and three pusher type furnaces. After discharging, the slab is rolled at the reversing roughing mill, typically in seven passes, in order to reduce its thickness from some 210 millimeters to the target thickness of e.g. 30 millimeters. After roughing, the rolling stock is called a transfer bar.

Next, the transfer bar goes through the finishing mill. In the case of the most recent datasets, a coil box had been added on the line right before the finishing mill. The coil box makes rolling of bigger coils possible and enables better control of the uniform temperature of the transfer bar. After finishing, the rolling stock, which is now called the hot strip, enters the cooling area where the targeted temperature profile of the strip is achieved. Finally, the strip is coiled at the down-coiler.

During the rolling process, many measurements are made and part of the meter positions are shown in Fig. 7. Temperatures are measured at several places, from the furnace up to the coiler, whereas width is measured at the rougher and after finishing. Thickness, profile, flatness, and the speed of the strip are measured after the finishing line. Also, the time stamps and the durations of the various process stages are recorded. From these measurements, the process

control system calculates the characteristics of the rolling stock: mean values, variations, segmented values, deviations from the targets, classification, etc., which are subsequently stored in the database.

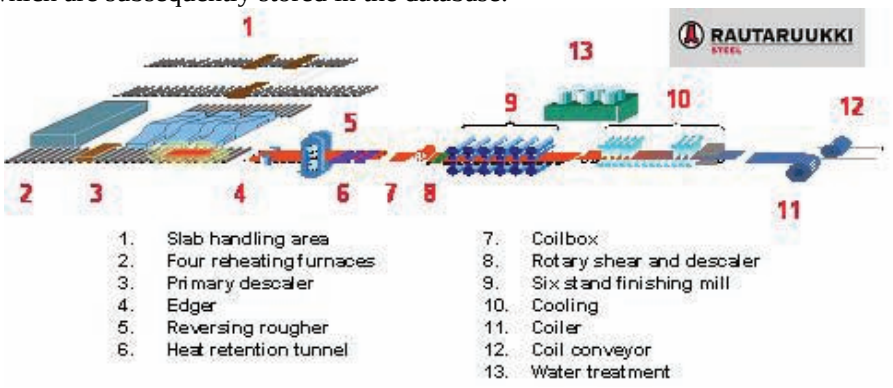

Fig. 6. Hot strip mill of Ruukki.

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 and uniform temperature during the process is a particularly challenging task. At any moment during the rolling, the automation system or the rolling mill operator may store a strip-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 thesis, these products are collectively termed retained samples.

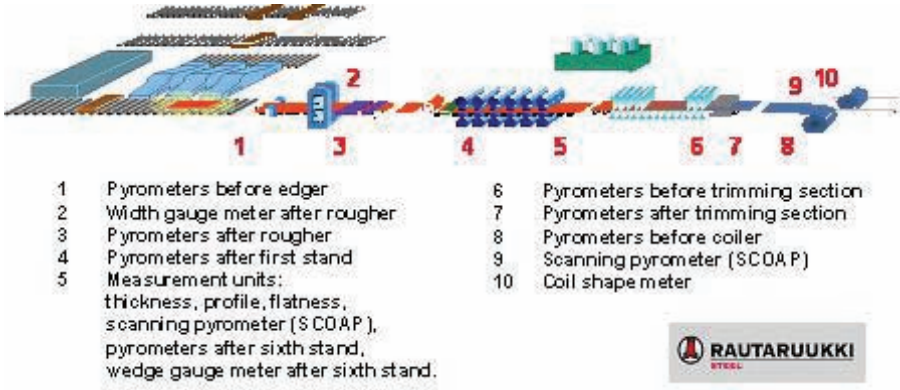

Fig. 7. Some meter positions at hot strip mill of Ruukki.

4.1.1 Original datasets

In the first part of the study, three different datasets were used for the development of the conditional probabilities method. These datasets are called collectively the original datasets. Below, is the short description of the datasets and of how they were utilized (see also Table 1).

The first dataset included 238 features with 4293 samples (see Appendix 1 for feature listing). 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 of this analysis were then presented to the rolling experts, who decided which of the features were to be included in the second dataset.

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. The dataset included 28738 samples. There were 9010 retained samples, i.e. 31.4% of all samples. The methods of linear correlation study, SOMs and k-means clustering were applied to this dataset, which resulted in conditional probabilities for retention codes.

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

Table 1. The numerical data of the three original datasets.

Dataset	Number of features	Total number of samples	Number of retained samples	Percentage of retained samples
First	238	4293	1086	25.3%
Second	41	28738	9010	31.4%
Third	12	33524	8743	26.1%

4.1.2 Most recent datasets

The results that were gained with the original datasets used for the development of the method suggested that *the thickness of the slab* was a dominant feature of the data. Moreover, some of the conditional probabilities obtained could not be verified (see Chapter 5). Hence, once again, a new dataset was gathered from

Ruukki. This time, the first dataset (including 51 features and 44965 samples) was further divided into twelve different subsets by using the retention codes and the values of the target thickness for hot strip rolling. In the group of the thin strips, the target thickness for hot strip rolling was less than or equal to 3.5 millimeters; in the group of the medium thick strips, the target thickness was more than 3.5 millimeters and less than or equal to 6.5 millimeters; and in the group of the thick strips, the target thickness was more than 6.5 millimeters. From the retention codes, four groups were selected: temperature related codes, telescope related codes, dimension related codes, and torn tail related codes.

The temperature related retention codes were: too low rolling temperature (9 samples), too high rolling temperature (1946 samples), and wrong coiling temperature caused by the cooling system (3244 samples). The telescope related codes were: error with telescope of under 60 millimeters (157 samples) and retention caused by a telescope (578 samples). The dimensional retention codes were: too thin (962 samples), too thick (334 samples), too narrow (2999 samples), and too wide (484 samples). Finally, the torn tail related codes were: error with slightly torn tail (1 sample) and retention with largely torn tail (97 samples). All these limits and codes to be studied were set and chosen by the rolling experts. From the resulting twelve subsets, five included retention codes that represented less than 1% of the subset size. With such a small amount of retained samples, the basic statistical analysis could not find any significant features and, thus, these five sets had to be excluded from the further studies.

The numerical data of the valid seven subsets is presented in Table 2. The first column indicates what values of the retention codes and the target thicknesses were used to form the subset, the second column indicates the total number of samples in the corresponding subset, the third column indicates the number of retained samples in the corresponding subset, and the last column indicates the percentage of retained samples in the corresponding subset. From the values in the below table, it can be concluded that the hot strip rolling of thick strips is more difficult than rolling medium thick or thin strips, because in the two subsets of the thick strips, more than every fourth strip is retained in contrast to a maximum of every ninth strip in the other five subsets. Hence, in all future analysis concerning hot strip rolling, it should be at least noted, if the target thickness of the strip is likely to have a considerable impact on the results.

Table 2. The numerical data of the seven subsets.

Subset	Total number of samples	Number of retained samples	Percentage of retained samples
Temperature / Thin	18987	2103	11.1%
Temperature / Medium Thick	13758	1252	9.1%
Temperature / Thick	6004	1529	25.5%
Telescope / Thin	17533	649	3.7%
Dimension / Thin	18507	1623	8.8%
Dimension / Medium thick	14003	1497	10.7%
Dimension / Thick	5998	1523	25.4%

Even though all of these seven subsets are presented here for the purpose of understanding the affect of thickness, only the results concerning the temperature and telescope retentions are presented later in this thesis (see Chapter 5.2). The reason for this is that the results obtained for subsets with dimensional retentions have not yet been discussed with the rolling experts and, thus, they cannot be published at this moment.

4.2 Steel Dynamics Inc.

The continuous casting technology was developed already in 1960s and, thus, it is quite well studied. Nowadays, since a high percentage of the produced slabs is defect-free and need not to be conditioned, continuous casting enables direct connection between casters and hot rolling mills. Consequently, direct rolling reduces energy consumption and, consequently, the cost of energy (Wiesinger *et al.* 1985). Nevertheless, there is a continuous demand for improvements in product quality and this includes also higher standards of dimensional accuracy. One of the dimensional elements, or thickness profiles, is wedge (Ginzburg 1993, Mücke *et al.* 2002). The term wedge describes the workpiece profile asymmetry in qualitative terms. The drive side wedge is identified as a strip profile with center gauge being less than the drive side thickness and greater than the operator side thickness. As per the symbols from Fig. 8, the drive side wedge is present when td > tc > to. Correspondingly, the operator side wedge is identified as a strip profile with center gauge being less than the operator side thickness and greater than the drive side thickness. In other words, the operator side wedge is present when to > tc > td (see Fig. 8).

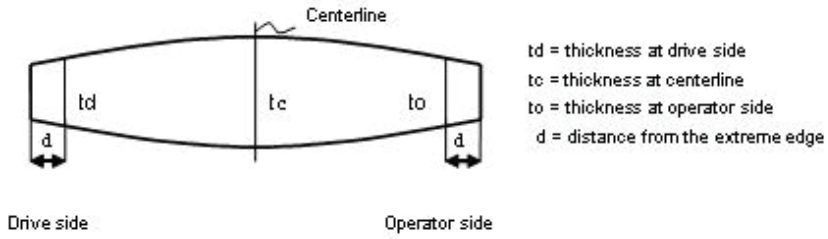

Fig. 8. Outline of a section across the width of the strip.

A small positive center crown is usually needed in order to achieve a good tracking of a strip (Mücke *et al.* 2002, Shiraishi *et al.* 1991). The center crown is defined as the difference between the center gauge and the edge thicknesses:

$$\text{center crown} = tc - (td + to)/2 \tag{15}$$

Even a small reduction in crown or a minuscule variation in the thickness profile of the incoming strip can easily cause, for example, a walking of the strip, which, in turn, can increase both wedge and camber (see Fig. 9). These two properties are considered as signs of poor quality. Other sources for wedge are casting and slitting operations. In addition, Masui *et al*. and Tarnopolskaya *et al*. (Masui *et al.* 2000, Tarnopolskaya *et al.* 2002) demonstrated that a wedge-shaped incoming strip can produce a walking of the strip, which, then, can increase wedge that produces incremental walking, and so forth. In their study of relation between camber and wedge (Shiraishi *et al.* 1991), Shiraishi *et al*. concluded that it is very difficult to reduce both camber and wedge at the same time during finishing rolling. As a result, they suggest that camber and wedge be mainly controlled during the roughing stage. Biggs *et al*. (Biggs *et al.* 2000) showed ways to minimize camber during the early reduction stages in a hot strip mill, but these do not necessarily reduce the wedge. By using finite element calculations, Loney *et al*. (Loney *et al.* 2002) were able to produce statistical models for camber, wedge and sideways movement. These models were evaluated with pilot plant trials and the models were found to be good. The features used were *strip width* (see Fig. 9), *reduction, positioning, tilt of the mill*, and *work roll shape*, which were all found to have a significant impact on all three events under study. In addition to these results, finite element simulations showed that an initial wedge of the slab can be decreased by rolling with an opposite wedge roll gap in the first pass of the roughing mill.

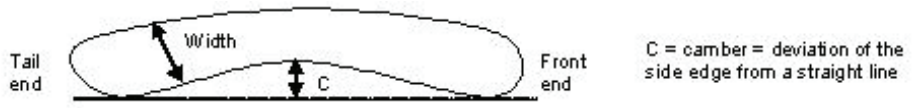

Fig. 9. View from above the strip.

At the beginning of 1990's, Louhenkilpi (Louhenkilpi 1990) stated that continuous casting had become the dominant process route in modern steel production, replacing ingot casting. He also pointed out that the major advantages of continuous casting compared with ingot casting were improvement of steel quality, better yield and savings of energy and manpower. As the process of hot strip rolling with continuous casting is dissimilar to a process of hot strip rolling with slab reheating furnaces, it was felt that these two cases would be different enough so that the case of continuous casting could be used as a verifying case for the developed method.

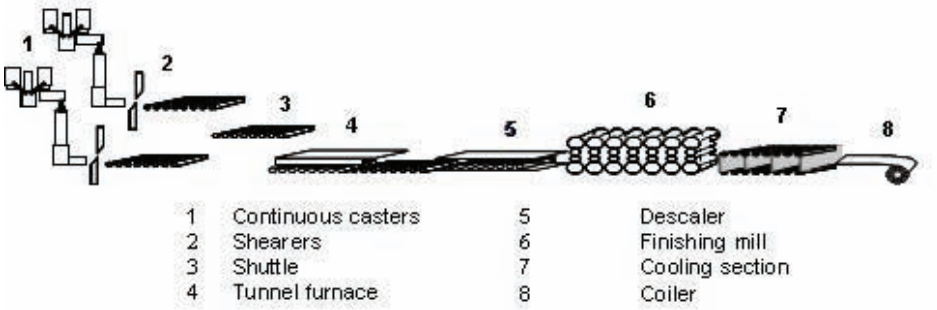

Fig. 10. The continuous casting mill of Steel Dynamics, Inc.

The data was gathered from the continuous casting mill of Steel Dynamics, Inc. in Butler, Indiana, USA. The mill uses the Compact Strip Production (CSP) thin-slab process to produce sheet steel as illustrated in Fig. 10. Before the hot strip rolling process, the 55-millimeter thick slabs are fabricated by two continuous casters, which are synchronized to produce slabs in alternate turns. Right after the slab is bent and straightened into horizontal level, it is mechanically sheared to ordered weight and is then moved to the rolling line by shuttle. At the beginning of the rolling, the slab is kept at the right temperature with a tunnel furnace, which, at the same time, provides a time buffer between the caster and the hot strip mill. Next, the slab goes through the descaler and the finishing mill, which has seven stands. Finally, the strip is cooled and coiled.

Unlike a traditional flat-rolled mill, this mill does not use a slab yard to store intermediate products and, therefore, does not require reheating furnaces. The CSP process has lower cost and is more efficient than the traditional production process, but its downside is that it lacks buffers in the production process. If one part of the CSP mill breaks down, the entire mill has to be shut down until the mill is repaired. However, it is not costless to carry out a slab inventory.

The original dataset included 49 features and 39491 samples. After some preliminary cleaning of the given dataset, 38 features remained. At the beginning of the study, the dataset was randomly divided into two parts, namely, development and evaluation datasets. In order to verify the similarity of these two datasets, comparisons between means, variances, and value ranges of all features were made. One of the features included *the measured wedge of the strip* and this feature was later used as a classifying one. After the division of the dataset, some more samples were cleaned from the two newly formed datasets. As a result, the development and the evaluation datasets included 19486 and 19496 samples, respectively. For the purpose of the classification, the feature of measured wedge was converted into a class feature so that the values within two times the deviance of wedge from the average of wedge were replaced with value zero and the rest of the values were replaced with value one. In other words, the values within

average of wedge ± 2 * deviance of wedge

were replaced with value zero and the values outside these two limits were replaced with value one. The numeric value for the lower limit is -1.26 mils (1 mil = 0.001 inches) and for the upper limit 1.21 mils. As mentioned earlier, the value of wedge is usually within the acceptable limits and, hence, there were no predetermined limits for this thesis. The interval that was used here was selected together with the rolling experts.

Having done the conversion of the measured wedge, the development dataset included 1054 samples that belonged to class one, i.e. 5.4% of the samples had a high value of wedge. For the evaluation dataset, the corresponding number was 1047 samples, i.e. 5.4% of the samples belonged to class one and, thus, had a high value of wedge. Table 3 includes the numerical data of the two datasets: the first column indicates the names (and the purposes) of the datasets, the number of used features is in the second column, the third column shows the quantities of samples in each dataset, the fourth column gives the quantities of samples with large measured wedge per dataset, and the last column gives the percentages of samples with large measured wedge within each dataset.

Table 3. The numerical data of the Steel Dynamics datasets.

Dataset	Number of features	Total number of samples	Number of samples with large wedge	Percentage of samples with large wedge
Development	37	19486	1054	5.4%
Evaluation	37	19496	1047	5.4%

4.3 Discussion

In order to evaluate the usability of the developed conditional probabilities method, it was thought necessary to have two different targets of application. By this, it was possible to show that the method was not suitable for only one type of problem, but that it could be used also in another environment.

Although the two targets of application in this thesis are both cases from steel manufacturing by hot strip rolling, they are, by nature, very different. In Ruukki process, it is possible to measure (and calculate) more features that are related to the process before it actually starts, because the used slabs are taken from the storage instead of being casted on-the-fly as it is in the case of Steel Dynamics. Having a greater amount of knowledge beforehand makes it possible to plan the act of rolling in more detail. The result of this can be seen, for example, when comparing the quantities of features in the original datasets obtained from the two factories (238 vs. 49).

In addition to the obvious differences in the actual rolling processes, the study had very different goals for the two cases: in the case of Ruukki, the goal was to discover the possible patterns behind the occurrence and the absence of retentions, which are sums of many different features of the finished products, whereas, in the case of Steel Dynamics, the goal was to discover which features affect the behavior of another feature i.e. the wedge of the strip.

Now, given these dissimilarities between the two cases, the condition of having two separate study environments was seen to be fulfilled. Consequently, the functionality of the developed method could be verified reliably.

5 Conditional probabilities for retentions at Ruukki

The results of the retentions at Ruukki were divided into two parts according to the datasets that were used: the original datasets and the most recent datasets. These datasets are described in detail in Chapter 4.1.

5.1 Original datasets

The group of original datasets includes three different datasets that were used for the development of the conditional probabilities method (see Chapter 4.1.1). These are further named as first, second, and third dataset in the order of their utilization. The results presented in this chapter are derived from these datasets. Fig. 11 shows the progress in the quantities of features and components for these three datasets. In this figure, the values beside the arrows indicate the numbers of features or components to be included in the next stage of the process. In addition to these features or components, there is always at least one labeling feature alongside that is not included in the given amounts. One exception exists, however, with the number of original features (238) that includes multiple labeling features.

5.1.1 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 then provided detailed information about each feature, e.g. how the features were formed or measured and if the values were reliable. Then, the experts decided on the contents of the second dataset, which now included features like *slab's material group, slab's quality code, slab's dimensions, measured temperature at roughing mill, calculated tensile strength,* and *slab's surface quality.*

The analysis of the second dataset included 41 features. Thirty-one of these were found to be significant in the sense that the retentions were not evenly distributed over the values of the feature. 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 second dataset was then used for deriving the conditional probabilities.

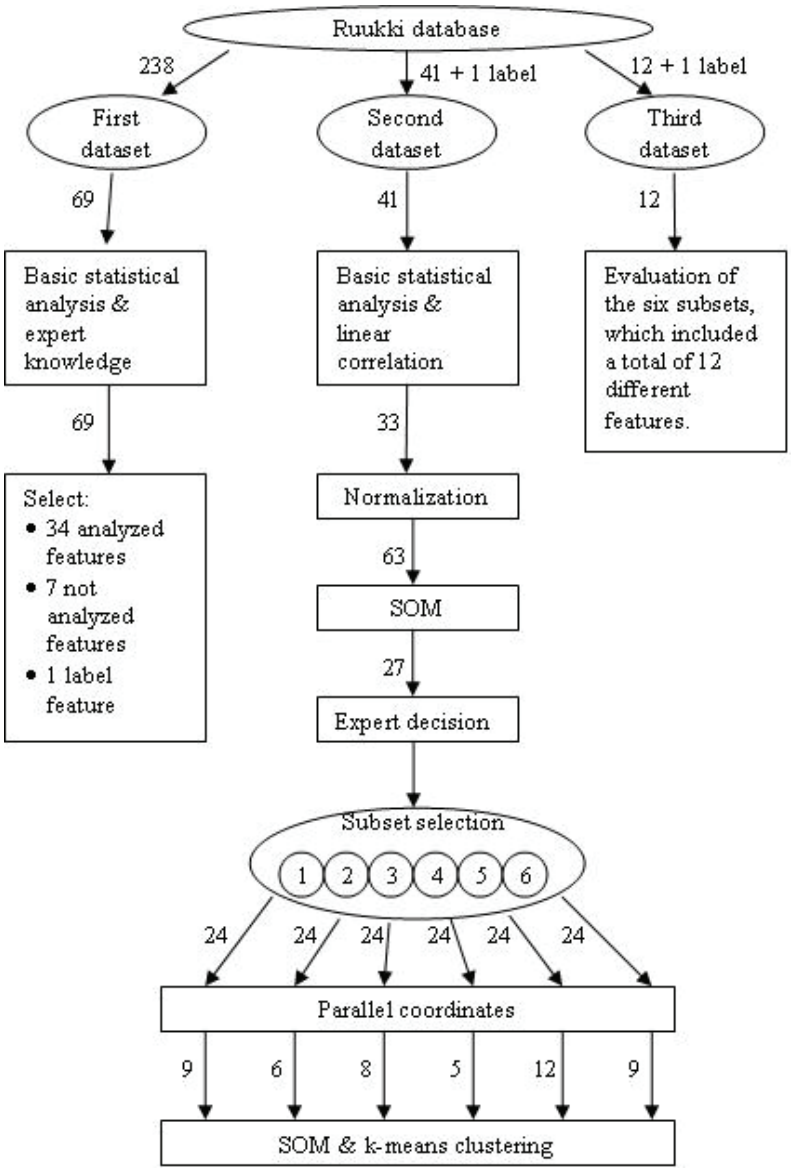

Fig. 11. The amounts of features and components in each Ruukki dataset.

When the statistics of these 41 features were studied more closely, it gave an idea of what should be the limits for finding the significant features according to the rules in Chapter 3.1.1. The specific limits were then selected by using the knowledge gained from the statistical studies and although they were challenged throughout this thesis, they seemed to work appropriately in all cases. In consequence, it is not argued that these exact limits should be used in all situations, but, rather, that they could always be used as a starting point to find the best limits for each particular case.

The final analysis was done to the third dataset with twelve features. This dataset was used to evaluate the conditional probabilities that had been derived by using the second dataset. The twelve 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, statistical comparisons were made between the twelve features in both of them. These comparisons included the mean, standard deviation, and value range for both the whole dataset and the set that only included the samples with retentions. The hot strip rolling process, from which the features were measured, was altered between the two time spans of the measurements. A transfer bar coiler (coil box) 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 retentions.

5.1.2 Linear correlation and SOM

The use of SOM requires an exclusion of the repetitive features from the dataset. At this point, there were 29 continuous and 12 class features in the second dataset. Thus, the linear correlation analysis was done to the continuous features, leading to the exclusion of eight features. In addition to the labeling feature (that included the retention codes), the resulting dataset now included 21 continuous and 12 class features and, hence, the size of the dataset (42) was 14% of the original size of 238 features (that included multiple labeling features).

When applying SOM, the features have to be normalized from 0 to 1 or from –1 to 1. In this study, the former scale was selected. The normalization of the features used in this study (33) 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 further analysis because they seemed to have an effect on retention distribution in the

clusters. Fig. 12 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 darker areas. In Fig. 12, the areas separated with a solid line included extensively more samples with retention code than the areas marked with a dotted line.

A second SOM was calculated with the selected 27 components. The visual inspection of the resulting component planes did not reveal any need to reduce the number of the components, but, in order to obtain predictive rules of retention appearance, the three components measured after the finishing mill had to be excluded from further analysis.

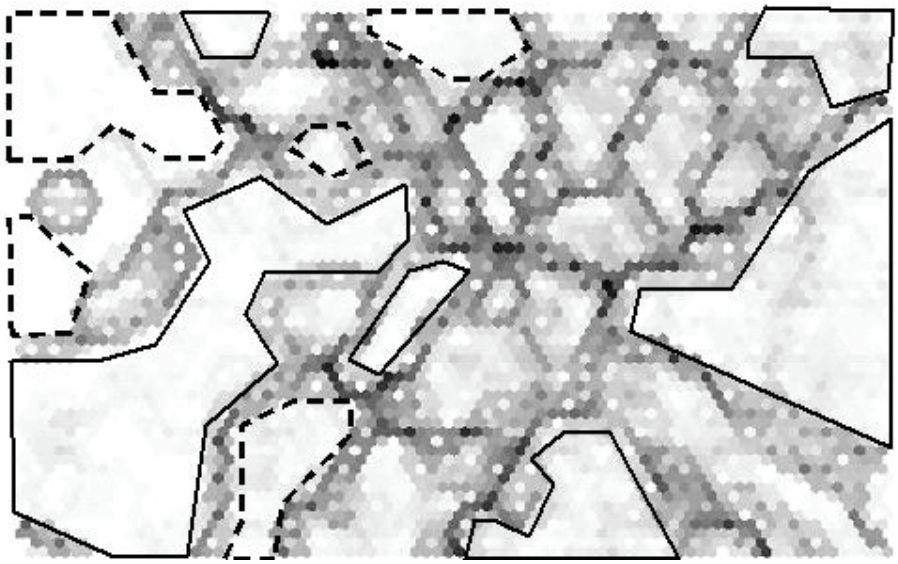

Fig. 12. The SOM of 27 components.

5.1.3 Parallel coordinates

At this point, in order to find out the explanations for different retention types, six smaller development datasets were selected according to the most common retention codes (see Fig. 11 and Table 4). 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 any retention code and samples that

had a certain type of retention code. Table 4 shows some numerical data of these six smaller datasets: the first column shows the retention type corresponding to each dataset, the second column indicates the quantity of components, the third column lists the numbers of samples per dataset, and the last column gives the percentages of retained samples within the datasets.

When looking at the names of the retention types, they indicate that there exist numerical limits for labeling a sample with a retention code. This is true to some extent, as there are guidelines for these limits, but, in reality, the final decision of assigning a retention code is made by a rolling expert based on his or her expertise.

Next, the six datasets were 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. Fig. 13 presents the parallel coordinates for the first set, which includes both successful samples and samples having a too high rolling temperature. 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.

Table 4. Some numerical data of the six smaller development datasets.

Subset	Number of components	Total number of samples	Percentage of retained samples
Dataset 1: Rolling temperature too high	24	21002	6.2%
Dataset 2: Coiling temperature close to the limits	24	23366	15.7%
Dataset 3: Too thin	24	19972	1.4%
Dataset 4: Too narrow	24	20734	5.0%
Dataset 5: Too wide	24	19855	0.8%
Dataset 6: Torn tail end	24	21192	7.1%

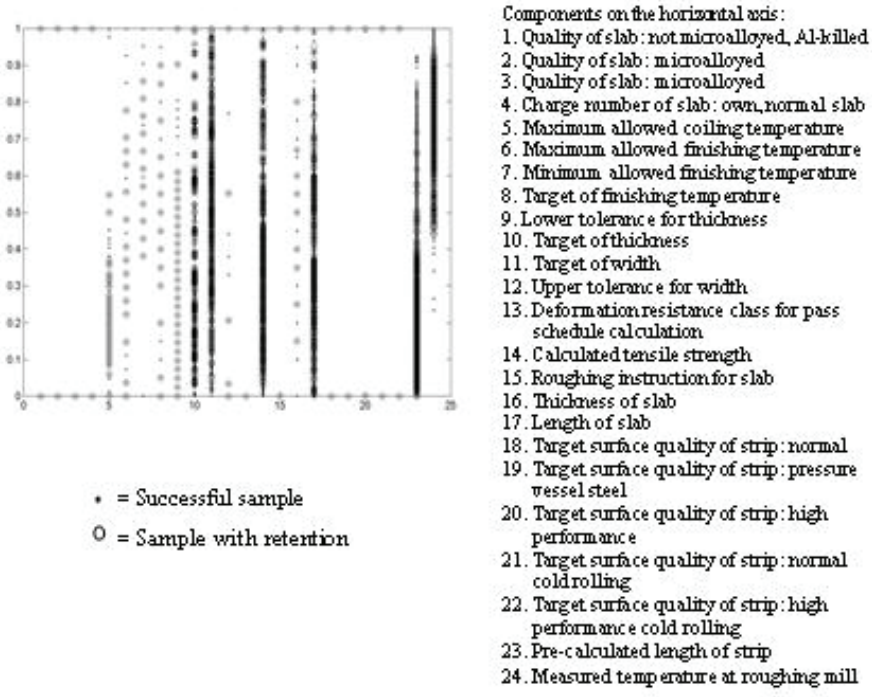

Fig. 13. Parallel coordinates for the first smaller dataset: components are aligned on the horizontal axis and their normalized values are on the vertical axis.

In Fig. 13, for the high values of 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 components 16 and 24. What is more, components 9 and 12 seem to have some differences between the distributions of the successful samples and the samples with retention codes. Hence, components 5, 6, 7, 8, 9, 12, 16, 23 and 24 were selected for the next stage of the study. These components correspond to the following features: *maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, upper tolerance for width, thickness of slab, pre-calculated length of strip,* and *measured temperature at roughing mill.* Obviously, having higher values for minimum, maximum, and target temperatures would not produce a retention due to too high rolling temperature, but the interpretation of the other components is not as straightforward. For example, thickness of slab (component 16) has,

throughout the range, plenty of values that do not correlate with the related retention, but also several values that do. The remaining sets were also studied in a similar manner. In addition to the nine features selected above, the features *length of slab, target of width,* and *calculated tensile strength* were proved to be significant in explaining the differences between successful and retained samples. The selected components for each of the six smaller dataset are listed in Table 5.

Table 5. The selected components of the six datasets after the analysis with parallel coordinates.

Dataset	Selected components (number)
Dataset 1: Rolling temperature too high	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, upper tolerance for width, thickness of slab, pre-calculated length of strip, measured temperature at the roughing mill (9)
Dataset 2: Coiling temperature close to the limits	minimum allowed finishing temperature, target of finishing temperature, thickness of slab, length of slab, pre-calculated length of strip, measured temperature at the roughing mill (6)
Dataset 3: Too thin	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, thickness of slab, length of slab, measured temperature at the roughing mill (8)
Dataset 4: Too narrow	minimum allowed finishing temperature, upper tolerance for width, thickness of slab, pre-calculated length of strip, measured temperature at the roughing mill (5)
Dataset 5: Too wide	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, target of width, upper tolerance for width, calculated tensile strength, thickness of slab, length of slab, pre-calculated length of strip, measured temperature at the roughing mill (12)
Dataset 6: Torn tail end	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, calculated tensile strength, thickness of slab, pre-calculated length of strip, measured temperature at the roughing mill (9)

5.1.4 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 (Davies & Bouldin 1979).

Fig. 14 and Fig. 15 present the self-organizing map with the clustering and the corresponding component planes for the dataset including retention due to a too high rolling temperature. In Fig. 14, the map includes 600 knots and the number in each knot represents the cluster that the knot belongs to. Every knot includes an undetermined quantity of samples that are very alike and the samples in a certain knot are represented with a 9-dimensional mean vector, which is visually expressed with component planes in Fig. 15. The colorbars beside each component plane represent the normalized values for each component. For example, the knot in the upper left corner has a mean vector where the seventh component, i.e. *the thickness of slab*, has a high value (= dark color), when, on the other hand, the knot in the lower right corner has a low value (= light color) for that same component. As it can be seen from Fig. 14, the clusters vary in shape and in size and they can be unconnected like, for example, cluster number 3. Actually, when interpreting the map, it should be remembered that the location of a knot bears no meaning, as is the case with the size and the shape of the clusters. In fact, the nature of the SOM program is the only reason why similar knots appear mostly beside each other.

In order to interpret the results, just by visually comparing these two figures, it may be concluded that the components *minimum allowed finishing temperature, upper tolerance for width* and *thickness of slab* play an important role in dividing the dataset into clusters. To make this more visible, boundaries for clusters 3 and 5 are also drawn into each component plane. For example, the higher values of the component thickness of slab (lower left corner in Fig. 15) are almost totally separated from the lower values, and they seem to fall into only one cluster (numbered as 5 in Fig. 14).

This type of visual inspection gives an idea on how these components relate to each other, but, in order to find out the relations between the retentions and the components, some numeral values are required. In the present case, the clusters efficiently separated the parts of the map that only had successful samples from the parts that had multiple samples with retention. For example, cluster number 5 included 3484 successful samples and none of the samples with retention. On the other hand, cluster number 7 included 862 of the successful samples and 265 samples with retention.

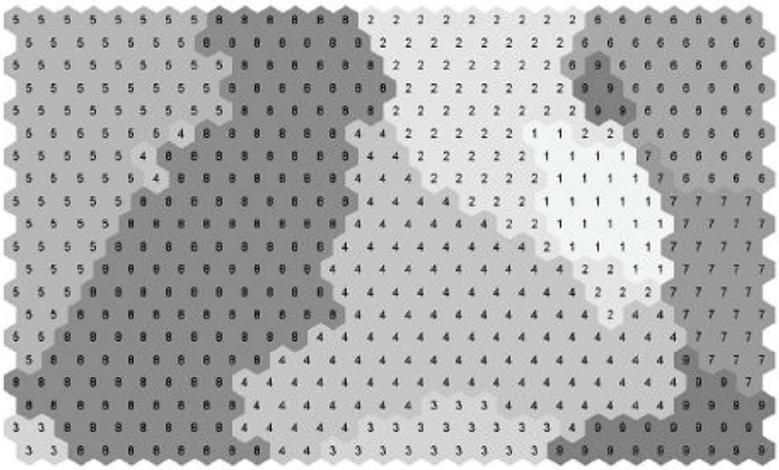

Fig. 14. The clustered self-organizing map for the dataset including retention due to a too high rolling temperature; incorporates nine clusters, the number in each knot representing the cluster that the knot belongs to.

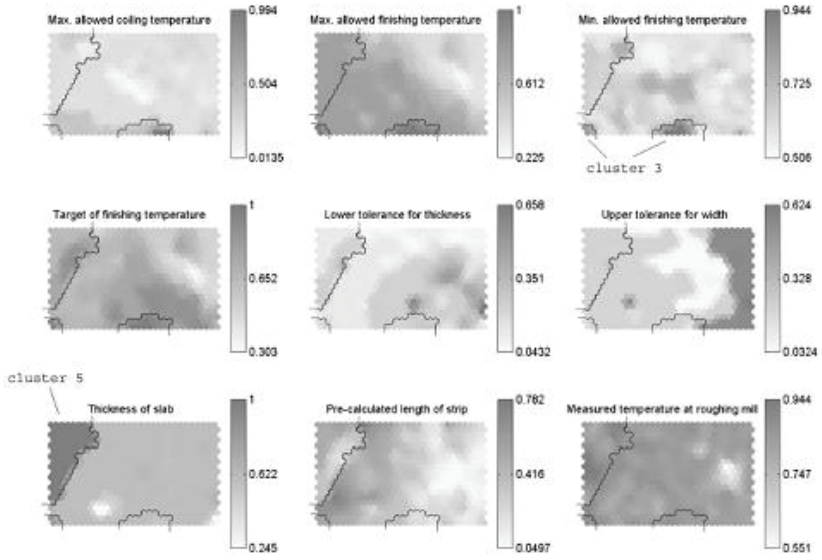

Fig. 15. The component planes and the colorbars of the normalized component values for the dataset including retention due to a too high rolling temperature; boundaries for clusters 3 and 5 are drawn into each component plane.

Hence, for cluster 5, the probability of having a retention 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 for a sample that resembles more the ones in cluster 5 than those in cluster 7. After studying the component values corresponding to the clustering, the conditional probabilities were derived for each cluster.

In the example presented above, the conditional probability for cluster number 5 is:

Probability for having a retention due to a too high rolling temperature is 0.000 with conditions:

	0	≤	*maximum allowed coiling temperature*	≤ 0.42
and	0.70	≤	*maximum allowed finishing temperature*	≤ 0.93
and	0.57	≤	*minimum allowed finishing temperature*	≤ 0.76
and	0.60	≤	*target of finishing temperature*	≤ 0.90
and	0.02	≤	*lower tolerance for thickness*	≤ 0.71
and	0.03	≤	*upper tolerance for width*	≤ 0.55
and	0.80	≤	*thickness of slab*	≤ 1
and	0.05	≤	*pre-calculated length of strip*	≤ 0.62
and	0.56	≤	*measured temperature at the roughing mill*	≤ 1.

Correspondingly, the conditional probability for cluster number 7 is:

Probability for having a retention is 0.235 with conditions:

	0.06	≤	*maximum allowed coiling temperature*	≤ 0.31
and	0.07	≤	*maximum allowed finishing temperature*	≤ 0.70
and	0.52	≤	*minimum allowed finishing temperature*	≤ 0.76
and	0.10	≤	*target of finishing temperature*	≤ 0.90
and	0.05	≤	*lower tolerance for thickness*	≤ 1
and	0.55	≤	*upper tolerance for width*	≤ 1
and	0.25	≤	*thickness of slab*	≤ 1
and	0.34	≤	*pre-calculated length of strip*	≤ 1
and	0.43	≤	*measured temperature at the roughing mill*	≤ 0.95.

As it can be seen, the conditions in these two clusters differ remarkably for components 2, 4, 6, 7, and 8. For example, component 2, which is *the maximum allowed finishing temperature*, has values greater than or equal to 0.7 in cluster

number 5 while, in cluster number 7, the values are smaller than or equal to 0.7. With this component, the value 0.7 corresponds to 940 degrees of Celsius. From the point of view of steel production, it is vital to have clusters with zero probability for having retention, because by looking at these clusters and their conditions, the rolling experts can plan their rolling campaigns in a way to reduce the probability of retentions.

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, and the samples were linked to the existing clusters. When comparing the proportional numbers of retentions 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. In addition to reliability, these sets had clearly different rolling conditions between the clusters and it was, thus, simple to trace the circumstances of having a good product and to separate these distinctly from the case of having a retained coil. Furthermore, it became easy to reveal the conditions that are causing the four types of retentions. These retentions are: coiling temperature close to the limits, too thin, too narrow, and torn tail end. The other two datasets, which included the retentions rolling temperature too high and too wide, had some changes in their distributions of proportional numbers of retentions and they had to be presented to the rolling experts before utilizing the results. As an utcome of the discussion with the rolling experts, it was suspected that the thickness of slab was somewhat dominating the results and that this influence should be eliminated from the further studies. Even knowing this and looking at the results of the basic statistical analysis would not have been sufficient in order to identify this type of domination.,which was only discovered with the results of the conditional probabilities. Thus, yet a different data was selected to be analyzed. The results of that study are presented in the next chapter.

5.2 Most recent datasets

After analyzing the results of the earlier studies together with the rolling experts, it was decided to gather up new data with 51 features and with retentions concerning temperatures, torn tails, telescopes, and dimensions. The decision of which features should be included was made together with the rolling experts and, here, the knowledge obtained with the previous results played an important role too. The selection of subsets and the data used in this section are described in

detail in Chapter 4.1.2. In addition to the 51 features, two new features were decided to be formed and, hence, the features now totaled 53. One of the self-formed features was *the tolerance of allowed coiling temperature*, which was calculated by subtracting *the minimum allowed coiling temperature* from *the maximum allowed coiling temperature*. The other was *the tolerance of allowed finishing temperature*, which was calculated in a similar manner. Some of the 53 features were used only with certain type of retentions, for example, *the analyzed carbon content* was used only with datasets including temperature retentions, while it was not used with dataset including telescope retentions. The following chapters include the results that were derived with four different datasets presented in the following two tables. Table 6 contains some numerical data of the four subsets: the number of the features, the total number of the samples and the percentage of the retained samples that were included in the preliminary subsets. Table 7 lists all the corresponding features for the three subsets with temperature retentions on one hand and for the subset with telescope retentions on the other hand.

Table 6. Some numerical data of the four subsets.

Subset	Number of features	Total number of samples	Percentage of retained samples
Temperature / Thin	37	18987	11.1%
Temperature / Medium Thick	37	13758	9.1%
Temperature / Thick	37	6004	25.5%
Telescope / Thin	46	17533	3.7%

At this time, the subsets were considered to be large enough to divide them into two, so that the conditional probabilities would be, at first, derived with the development subset and, then hopefully, verified with the evaluation subset. In order to do the division, a self-created Matlab script for random selection was used. The script was iterated so many times that both of the subsets included almost identical amount of retentions. The numerical data of these subsets is presented in Table 8, where the quantities and the percentages of samples and retentions are listed for both the development and the evaluation subset.

Table 7. The included features in two subsets.

Subset	Features (number)
Temperature	quality of slab, analyzed carbon content, charge number, planned use of coil box, realized use of coil box, length of head end with differing coiling temperature target, differing coiling temperature target at head end, tensile strength at coiling thickness and temperature, length of tail end with differing coiling temperature target, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed coiling temperature, tolerance of allowed coiling temperature, cooling strategy, maximum allowed finishing temperature, minimum allowed finishing temperature, target of allowed finishing temperature, tolerance of allowed finishing temperature, furnace number, target thickness for hot strip rolling, target profile, target width, deformation resistance class for pass schedule calculation, calculated tensile strength of strip, strip production class, production code, shift number, actual temperature of discharged slab, thickness of slab, duration of slab heating, length of slab, width of slab, weight of slab, pre-calculated length of strip, target thickness of transfer bar, target temperature of transfer bar after last pass (37)
Telescope	quality of slab, charge number, planned use of coil box, realized use of coil box, length of head end with differing coiling temperature target, differing coiling temperature target at head end, tensile strength at coiling thickness and temperature, calculated coil outer diameter, length of tail end with differing coiling temperature target, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed coiling temperature, tolerance of allowed coiling temperature, cooling strategy, maximum allowed finishing temperature, minimum allowed finishing temperature, target of allowed finishing temperature, tolerance of allowed finishing temperature, furnace number, lower tolerance for thickness, target thickness for hot strip rolling, upper tolerance for thickness, lower limit for tolerance and rejection of profile, target profile, upper limit for tolerance and rejection of profile, lower tolerance for width, target width, upper tolerance for width, calculated tensile strength of strip, strip production class, production code, shift number, actual temperature of discharged slab, thickness of slab, duration of slab heating, length of slab, width of slab, width of slab at the beginning of slab, weight of slab, width of slab at the end of slab, pre-calculated length of strip, target thickness of transfer bar, target temperature of transfer bar after last pass, target length of transfer bar, target width of transfer bar (46)

Table 8. Numerical data of the four development subsets and the four evaluation subsets.

Subset	Development subset		Evaluation subset	
	Number of samples	Number of samples with retention (%)	Number of samples	Number of samples with retention (%)
Temperature / Thin	9493	1051 (11.1%)	9494	1052 (11.1%)
Temperature / Medium Thick	6879	628 (9.1%)	6879	624 (9.1%)
Temperature / Thick	3002	762 (25.4%)	3002	767 (25.5%)
Telescope / Thin	8767	325 (3.7%)	8766	324 (3.7%)

5.2.1 Basic statistical analysis

At the beginning, the features in each of the four subsets were analyzed with basic statistical analysis. This was done for both the development subsets and the evaluation subsets in order to make sure that they were comparable.

Table 9. Features that were excluded from the four subsets according to basic statistical analysis.

Subset (number of remaining features)	Features that were excluded from further studies (number)
Temperature / Thin (33)	planned use of coil box, realized use of coil box, production code, shift number (4)
Temperature / Medium Thick (29)	planned use of coil box, length of head end with differing coiling temperature target, target profile, production code, shift number, actual temperature of discharged slab, thickness of slab, target temperature of transfer bar after last pass (8)
Temperature / Thick (31)	charge number, planned use of coil box, realized use of coil box, length of tail end with differing coiling temperature target, production code, thickness of slab (6)
Telescope / Thin (25)	charge number, planned use of coil box, realized use of coil box, calculated coil outer diameter, cooling strategy, minimum allowed finishing temperature, furnace number, lower limit for tolerance and rejection of profile, target profile, upper limit for tolerance and rejection of profile, lower tolerance for width, upper tolerance for width, strip production class, production code, shift number, actual temperature of discharged slab, thickness of slab, duration of slab heating, length of slab, target temperature of transfer bar after last pass, target length of transfer bar (21)

According to these results, some of the features where found to be insignificant and were, therefore, excluded from further studies. Table 9 lists the features that were excluded from each of the four subsets.

In the subset of telescope retention with thin strips, the number of features found to be insignificant (21) was not surprising in the context that the percentage of the retained samples was only 3.7%. Thus, it was unlikely that the percentage deviation of the retentions between the different values of the features exceed 4% (see the rules of categorizing features in the Chapter 3.1.1).

5.2.2 Linear correlation and SOM

After the basic statistical analysis, came the study of linear correlation. Analysis was carried out for the remaining continuous features, the excluded features being listed in Table 10. The last column of the table lists the features that were selected to be kept in the subsets in order to represent the excluded features.

Again, as was the case in basic statistical analysis, the correlation study was done for both the development and the evaluation subsets in order to verify the results. In all four cases, the study gave exactly the same results for the development and the evaluation subsets. Between the four subsets, the linear correlation values were quite similar too. In every subset, there was a very high correlation i.e. over 0.9 between the features *differing coiling temperature target at head end*, *differing coiling temperature target at tail end*, *maximum allowed coiling temperature*, *minimum allowed coiling temperature*, and *target of allowed coiling temperature*. From these, the feature *target of allowed coiling temperature* was selected to represent them all, because it had the highest average correlation with the other four features. In addition, for all four subsets, there was a high correlation between *target of allowed finishing temperature* and *maximum of allowed finishing temperature*, the latter of which was left in the subsets. It was indeed thought to be more adapted to the temperature related retention codes and, in the case of the telescope related codes, it had a higher sum of correlations with other features. In the case of the three subsets with temperature retention, a high correlation was also discovered between *target width* and *width of slab*. Fig. 16 presents the scatter plot between these two features in the subset with temperature retentions and thin strips. The horizontal axis includes the feature *target width* and *the width of slab* is on the vertical axis. As it can be seen from the figure, the linear correlation between these features is easily apparent. The feature *width of*

slab was selected to represent the two features, because the dimensions of slab are known earlier than the target values for the rolling.

Table 10. Features that were excluded from and kept in the four subsets according to linear correlation analysis.

Subset (number of remaining features)	Features that were excluded from further studies (number)	Features that were selected to present the excluded ones
Temperature / Thin (27)	differing coiling temperature target at head end, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed finishing temperature, target width (6)	target of allowed coiling temperature, maximum of allowed finishing temperature, and width of slab
Temperature / Medium Thick (23)	differing coiling temperature target at head end, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed finishing temperature, target width (6)	target of allowed coiling temperature, maximum of allowed finishing temperature, and width of slab
Temperature / Thick (25)	differing coiling temperature target at head end, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed finishing temperature, target width (6)	target of allowed coiling temperature, maximum of allowed finishing temperature, and width of slab
Telescope / Thin (16)	differing coiling temperature target at head end, differing coiling temperature target at tail end, maximum allowed coiling temperature, minimum allowed coiling temperature, target of allowed finishing temperature, target width, width of slab at the beginning of slab, width of slab at the end of slab, target width of transfer bar (9)	target of allowed coiling temperature, maximum of allowed finishing temperature, and width of slab

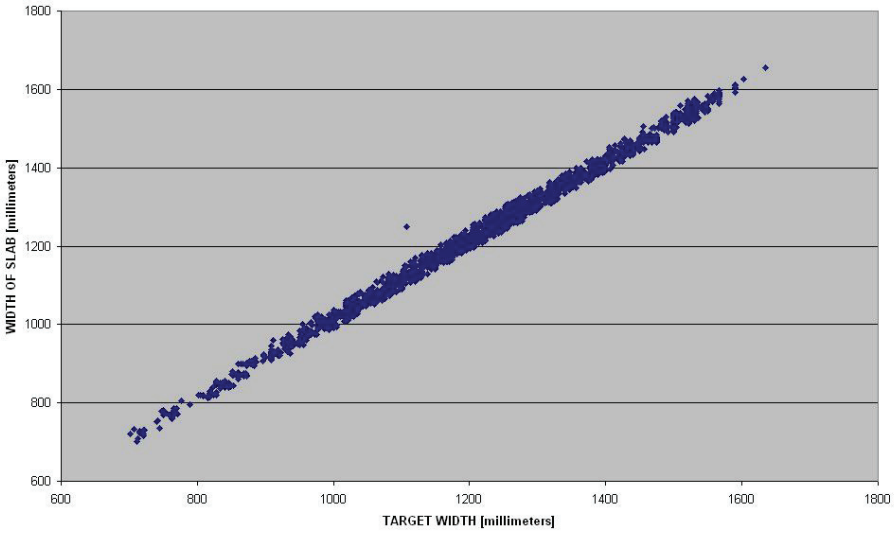

Fig. 16. The scatter plot between target width (horizontal axis) and width of slab (vertical axis) for the subset with temperature retentions and thin strips.

In the case of the subset with telescope related codes, there was a high correlation between *target width*, *width of slab*, *width of slab at the beginning of slab*, *width of slab at the end of slab*, and *target width of transfer bar*. From these five features *width of slab* was selected to represent all of them, because it had the highest average correlation with the other four features.

As the objective of using linear correlation was to remove some redundancy from the data, the limit for exclusion had to be carefully considered. For this purpose, different scatter plots were drawn. Fig. 17 shows the scatter plot between the features *tensile strength at coiling thickness and temperature* and *calculated tensile strength of strip* in the subset with temperature retentions and thin strips. The value for correlation between these two features was 0.8248, which appeared to be quite high and, thus, suggested that there was redundant information in them. But, by looking at the scatter plot, it can readily be seen that this was not the case this time. There is clearly a strong linear correlation between these two features, but the information in them is not redundant. Hence, these two were selected to be kept in the subset and the limit for the exclusion was set higher.

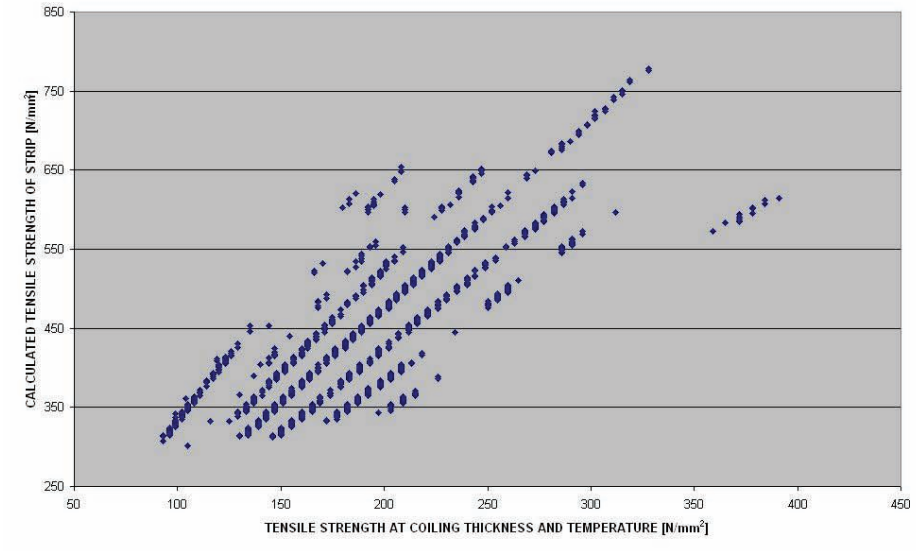

Fig. 17. The scatter plot between tensile strength at coiling thickness and temperature (horizontal axis) and calculated tensile strength of strip (vertical axis) from the subset with temperature retentions and thin strips.

In Chapter 3.1.2, it was stated that linearly uncorrelated features are not necessarily independent. A good example of this is presented in Fig. 18, which shows the scatter plot between the features *analyzed carbon content* and *length of head end with differing coiling temperature target* in the subset with temperature retentions and thin strips. The value for linear correlation between these two features was 0.1024, which appeared to be quite low and, thus, suggested that these features are uncorrelated. However, by looking at the scatter plot it can be seen that the data points are not randomly distributed in the plot, but, instead, are somewhat organized in multiple clusters. Hence, there is a dependency between these two features even though they are linearly uncorrelated.

All the remaining features were then normalized to have values within [0,1], so, that the self-organizing maps could be calculated. The varying number of the resulting components depends on the nature of the class features that are present in the data. All the methods that were used for achieving the normalization are presented in Chapter 3.4. The table below (Table 11) presents the number of features and components for each of the four subsets.

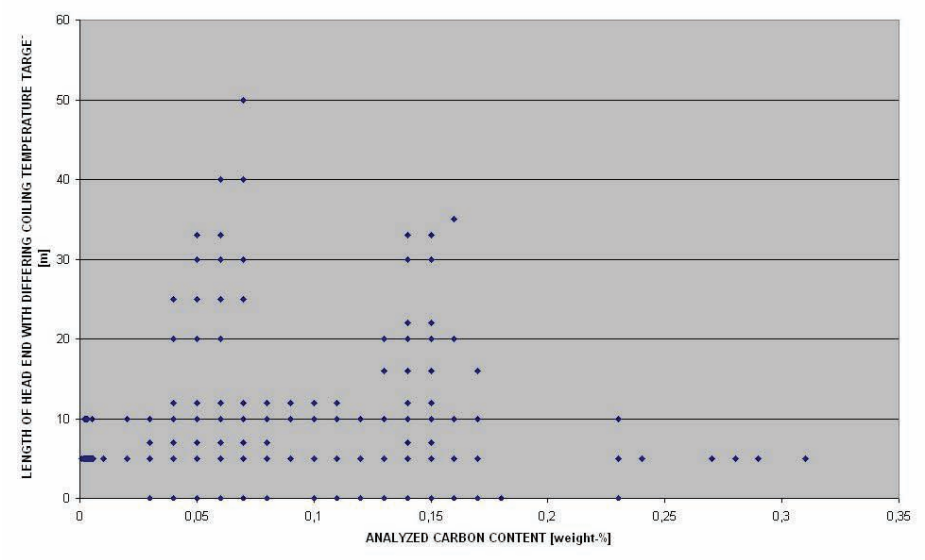

Fig. 18. The scatter plot between analyzed carbon content (horizontal axis) and length of head end with differing coiling temperature target (vertical axis) from the subset with temperature retentions and thin strips.

Table 11. The number of features and components in the four subsets.

Subset	Number of features before normalization	Number of components after normalization
Temperature / Thin	27	60
Temperature / Medium Thick	23	55
Temperature / Thick	25	62
Telescope / Thin	16	21

From this point forward, the evaluation subsets were put aside and the development subsets were used to do the calculations. The resulting SOMs revealed numerous clusters and, on the basis of a visual inspection of the corresponding component planes, some of the components were excluded from the further studies, because they did not seem to affect the retention distribution. Table 12 lists the excluded components.

Table 12. The components that were excluded from the further studies according to visual inspection of SOMs.

Subset (number of remaining components)	Components that were excluded from further studies (number)
Temperature / Thin (53)	deformation resistance class for pass schedule calculation = 10, deformation resistance class for pass schedule calculation = 11, deformation resistance class for pass schedule calculation = 13, strip production class = 6, strip production class = 8, strip production class = 9, weight of slab (7)
Temperature / Medium Thick (42)	quality of slab = 6, cooling strategy = 1, cooling strategy = 3, furnace number = 1, furnace number = 2, furnace number = 3, target thickness for hot strip rolling, deformation resistance class for pass schedule calculation = 3, deformation resistance class for pass schedule calculation = 4, deformation resistance class for pass schedule calculation = 7, strip production class = 3, strip production class = 5, strip production class = B (13)
Temperature / Thick (47)	quality of slab = 2, , cooling strategy = 1, furnace number = 1, furnace number = 2, furnace number = 3, furnace number = 4, target profile = 3, strip production class = 5, shift number = 1, shift number = 2, shift number = 3, shift number = 4, shift number = 5, duration of slab heating, target thickness of transfer bar (15)
Telescope / Thin (18)	quality of slab = 1, quality of slab = 2, quality of slab = 4 (3)

5.2.3 Parallel coordinates

The four subsets were then visualized by using parallel coordinates. The resulting figures were analyzed according to the manner explained in Chapter 5.1.3. Table 13 lists the remaining components for the four subsets after the analysis with parallel coordinates. With regards to the three subsets with temperature related retentions, there are quite many components that are important, and therefore, for example, *analyzed carbon content, tensile strength at coiling thickness and temperature, maximum allowed finishing temperature, calculated tensile strength of strip,* and *length of slab* were all included in the three subsets. But what is more interesting is that there are some differences between the subsets. For example, the components *target thickness for hot strip rolling* and *actual temperature of discharged slab* were not included in the subset with medium thick strips, while they were present in the other two subsets. The reason for this might be that it is much more demanding to make thick and thin strips as opposed to medium thick

strips and this difficulty also includes the notion that the temperature requirements are stricter with the strips that are rolled to be extremely thick.

Table 13. The remaining components after the analysis with parallel coordinates.

Subset (number of remaining components)	Remaining components after the analysis with parallel coordinates
Temperature / Thin (18)	analyzed carbon content, length of head end with differing coiling temperature target, tensile strength at coiling thickness and temperature, length of tail end with differing coiling temperature target, tolerance of allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, tolerance of allowed finishing temperature, target thickness for hot strip rolling, calculated tensile strength of strip, strip production class = A, actual temperature of discharged slab, thickness of slab, duration of slab heating, length of slab, width of slab, pre-calculated length of strip, target thickness of transfer bar
Temperature / Medium Thick (14 / 16)	analyzed carbon content, tensile strength at coiling thickness and temperature, length of tail end with differing coiling temperature target, target of allowed coiling temperature, tolerance of allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, calculated tensile strength of strip, duration of slab heating, length of slab, width of slab, weight of slab, pre-calculated length of strip, target thickness of transfer bar, **reinserted components:** strip production class = A, strip production class = B
Temperature / Thick (16)	analyzed carbon content, length of head end with differing coiling temperature target, tensile strength at coiling thickness and temperature, target of allowed coiling temperature, tolerance of allowed coiling temperature, maximum allowed finishing temperature, tolerance of allowed finishing temperature, target thickness for hot strip rolling, calculated tensile strength of strip, strip production class = B, actual temperature of discharged slab, length of slab, width of slab, weight of slab, pre-calculated length of strip, target temperature of transfer bar after last pass
Telescope / Thin (16)	quality of slab = 6, length of head end with differing coiling temperature target, tensile strength at coiling thickness and temperature, length of tail end with differing coiling temperature target, target of allowed coiling temperature, tolerance of allowed coiling temperature, maximum allowed finishing temperature, tolerance of allowed finishing temperature, lower tolerance for thickness, target thickness for hot strip rolling, upper tolerance for thickness, calculated tensile strength of strip, width of slab, weight of slab, pre-calculated length of strip, target thickness of transfer bar

The most interesting component seemed to be *strip production class*, which had one specific value in the subset of thin strip, no influence on the subset of medium thick strips, and another specific value in the subset of thick strips. After reviewing the results of statistical analysis concerning the subset of medium thick strips, it was decided to reinsert into the subset the components *strip production class equal to 'A'* and *strip production class equal to 'B'*. At the following stages, this subset was then treated as two separate cases: one subset not including the components of strip production class and the other containing them. Hence, the total amount of subsets to be handled at the last stage of the study was five instead of four.

5.2.4 k-means clustering

After the parallel coordinates, the remaining components were used to calculate the SOM, which was then clustered with k-means clustering. As results of these clusterings, the conditional probabilities were formed. With four of the five subsets used in this thesis, the best clustering was achieved with 20 to 25 clusters. This result implies that even though the subsets were products of careful selection of specific type of strips, there were still various process set-ups instead of just one, which could produce either wanted or unwanted products. Thus, it is just another proof that the process of hot strip rolling is a very complex combination of different factors. Yet, there was one exception with the subset of temperature retentions and medium thick strips including the two components of strip production class (see Chapter 5.2.3). With this subset the best clustering was achieved with four clusters. When the conditional probabilities for these four clusters were inspected more carefully, it became clear that the most influencing components, with regards to the SOM and the clustering, were the two strip production class components. In both the development and the evaluation subsets, there were less than two percent of such strips that belonged to these two specific steel grade classes 'A' and 'B', but, what is particularly interesting is that, jointly, there was only one retained strip. This is very good news because even though the strips belonging to these steel grade classes are rarely manufactured, they are almost always successful.

At the end of the study, the conditional probabilities were tested with the evaluation subsets (Table 14). At this point, the conditional probabilities of the subset with temperature retentions and medium thick strips, which did not include the two strip production class components, could not be verified sufficiently

enough, because there were some differences on how the samples from development and evaluation subsets were distributed in the clusters. This was the case especially with the retained samples. Hence, the results concerning this data are not reliable and are therefore not presented here. Instead, the conditional probabilities of the subset with temperature retentions and medium thick strips including the two components of strip production class were successfully verified with the evaluation subset. It is possible that by removing the samples that belong to the specific steel grades ‘A’ and ‘B’, the results obtained with the subset in question could be verified without the two components of strip production class, but this has not yet been studied more closely. The idea here is that the production class feature might have been influencing the calculation of SOM so strongly that the effects of the other features were lessened and, by removing this feature, the more subtle effects might be more easily detected. If this was the case, the results of clustering would be more likely succesfully verified. It is also a fact that the feature of strip production class has complex correlations with many other features, because the production classification incorporates information about, for example, the strengths, the risks, and the temperature regions of the slab to be rolled.

With the other three subsets (i.e. the subsets including temperature retentions and thin strips, temperature retentions and thick strip, and telescope related codes and thin strips), the conditional probabilities were successfully verified with the evaluation subsets. Table 14 summarizes the numbers of clusters in the best clusterings and the states of the verifications for each subset.

The next task was to convert the numerical limits of conditional probabilities into the form of more fluent text. In the case of the subset with temperature retentions and thin strips, there was one cluster that included all the samples belonging to the specific steel grade ‘A’ and none of the other samples. All 51 samples belonging to this cluster had been successfully rolled and, thus, this made the cluster a very interesting one. When this cluster was compared to the clusters that had a high percentage of retained samples, the following rule was derived: a successful coil is likely to be obtained, if a short, relatively narrow, and wide slab with high carbon content is processed to be thin, relatively short or long strip belonging to specific steel grade ‘A’ with 600-650 N/mm^2 tensile strength in conditions where the whole strip would have the same *target of allowed coiling temperature*, *the duration of slab heating* would be long, *the actual temperature of discharged slab* would be high, *the maximum allowed finishing temperature* would be high, *the minimum allowed finishing temperature* would be high or low,

the tolerance of allowed finishing temperature would be large, *the tolerance of allowed coiling temperature* would be large, and *the tensile strength at coiling thickness and temperature* would not be high. Corresponding sentences were derived also to express the probabilities of having a retained coil. All results were analyzed in a similar manner and were, then, presented to the rolling experts.

Table 14. Number of derived clusters after k-means clustering and the information about verification of the clustering.

Subset (number of remaining components)	Number of clusters in the best clustering	Verification of the clustering with the evaluation subset
Temperature / Thin (18)	20	Verified
Temperature / Medium Thick (14)	25	Not verified
Temperature / Medium Thick (16)	4	Verified
Temperature / Thick (16)	24	Verified
Telescope / Thin (16)	25	Verified

After discussing the results with the rolling experts, several conclusions were made. For example, the higher the target of the tensile strength of a strip, the slower the speed of the rolling process. And, when a rolling process is slow, the temperature of the strip drops more radically, because of the strong cooling effect, and, thus, the strip is more likely to have temperature retentions. Also, when the target thickness for hot strip rolling is over 6.5 millimeters (i.e. the strip is considered to be a thick strip) and the temperature tolerances are small, it makes the hot strip rolling a very demanding task. The reason behind this is that it is very difficult to get an even temperature distribution inside a thick strip. In general, these facts are already known by the experts and them being clearly present in the results, is a proof of the efficiency of the method that was used.

In addition to already known facts, there was one very intriguing result that could not be explained straightforwardly. When almost all rolling parameters are set to obtain successful strips, there is a striking difference in retention distribution between the strips with small carbon content (less than 0.23%) and the strips with higher carbon content (0.23-0.45%). From the group of strips with less carbon, more than one third (34.6%) of the strips were retained because of

the too high rolling temperature and, in the case of the strips with higher carbon content, only 0,4% of the strips were retained. According to present knowledge, the amount of carbon at this content level should not have a straightforward effect on the temperature behavior of the strip with the current temperatures and the speed of the rolling process. And still, even the mere statistical analysis gives a hint that the carbon content is a very important feature in this particular situation. Hence, it is suspected that there is a more complex effecter behind what is brought into light, in this case, with the carbon content. Nevertheless, this has been a new insight into the data for the rolling experts and studies to find the underlying reasons are still going on.

5.3 Discussion

In the case of the original datasets, the results of four out of six smaller datasets were verified with the evaluation data. For the retentions that were identified, it is now possible to see what kind of rolling conditions should be used to avoid specific types of retentions. The results also give information about the conditions that are likely to produce these retentions. When studying the reasons for failing to verify the conditional probabilities that were derived with the other two smaller datasets, two main factors were discovered: a change in the hot strip rolling process between the moments when the development and the evaluation datasets were retrieved and the thickness of slab.

The change in the hot strip rolling process was when a coil box was added in the process line before the finishing mill. This change affects the temperature distribution in the transfer bar. It is likely that this change had an influence on the occurrence of the retentions, especially, in the case of the rolling temperature being too high and this would be one of the reasons why the conditional probabilities that were derived including these specific retentions could not be verified.

When the results obtained with the original datasets were examined together with the rolling experts, it was found out that there was another factor that strongly affected the conditional probabilities: thickness of slab. And actually, this realization was one motive for the decision to study more recent data. Yet, it should be noted that the results that were achieved with the four smaller datasets are still valid in the sense that the presented method was able to find the most important features and their interactions behind the process under study. Basically, this example shows the iterative nature of the knowledge discovery

process, where the understanding of the problem is enhanced during the final stages and, then, the process can be reiterated more sophisticatedly.

In the case of the more recent datasets, the results showed that the method of conditional probabilities is capable of finding the complicated relations within the data, even though some components could not be handled in a straightforward manner with regards to the parallel coordinates. However, it should be remembered that this happened with only one of the datasets as opposed to the fact that most of the datasets were studied successfully according to the instructions. This issue was already discussed in Chapter 3.5 where it is noted that the use of this specific method should be addressed more carefully. The main proof that the derived method works properly lies on the fact that the developed conditional probabilities were verified in light of independent data.

Another factor proving the validity of the results was that they also included the facts that were already known by the rolling experts, i.e. more temperature related retentions occur when

1. the speed of the rolling process is slow or
2. the target thickness for strip is high (= over 6.5 millimeters).

In addition to these facts, the conditional probabilities included novel information on how the different rolling conditions affect the retention occurrence as the results indicated that the amount of carbon content (at the present content level) has an influence on the occurrence of the temperature retentions. This is valuable information for the experts and gives them yet another possibility to enhance the process of rolling.

6 Conditional probabilities for wedge formation at Steel Dynamics Inc.

The goal of the study with data from Steel Dynamics was twofold: from the application's point of view, the target was to discover new knowledge of how the different features affect wedge formation in a strip and, from the theory's point of view, the target was to show that the developed method of conditional probabilities would also work in another environment, where the quantity of features is smaller and the type of data is different.

At the beginning of the study, the provided dataset was randomly divided into two parts, namely, development and evaluation datasets. In order to verify the similarity of these two datasets, comparisons between means, variances, and value ranges of 38 features were made. One of the features included the measured wedge of the strip and this feature was later used as a classifying one (see Table 3 in Chapter 4.2). In addition, the datasets included features such as *the identification number of a caster*, *the grade of a strip* (based on qualification requirements by customers), *the average thickness of strip*, and *the height of the ridge on drive side*. In order to calculate the average thickness, the actual profile of the strip is measured right after the finishing mill and the nominal thickness is then determined by a non-linear curve fitting method. The ridge is defined as a deviation of the actual cross-profile from the fitted curve (see Fig. 19).

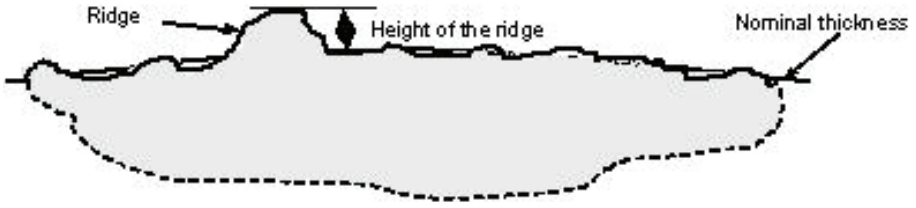

Fig. 19. Schematic presentation of nominal thickness and ridge.

Fig. 20 shows the progress in the quantities of features and components for the Steel Dynamics data and it is followed by the description of the feature selection and data evaluation process in this case.

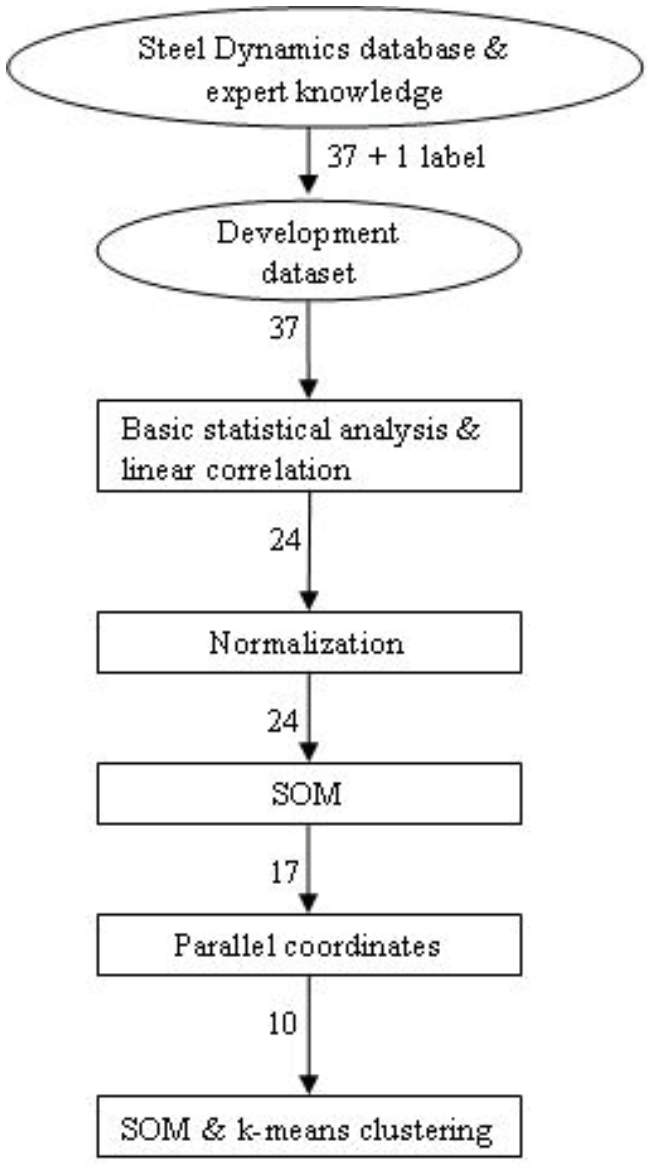

Fig. 20. The quantities of features and components in Steel Dynamics dataset.

First, the 37 features were analyzed with basic statistical analysis and, as a result, eleven features were excluded from further studies, because they were

insignificant in regard to the value of the wedge. These excluded features are listed in Table 15.

Table 15. Features that were excluded according to basic statistical analysis.

Features that were excluded from further studies (number)
identification number of caster, grade of strip, target value for center crown of strip, work roll wear of the first stand, work roll wear of the second stand, work roll wear of the third stand, work roll wear of the fourth stand, work roll wear of the fifth stand, work roll wear of the sixth stand, work roll wear of the seventh stand, roll change indicator concerning strip (11)

The list of excluded features does not really contain surprises, because, for example, *the work roll wear of a stand* is usually quite symmetrical and, thus, is not a candidate for producing asymmetrical wedge on a strip. The underlying explanation here is that the work roll wear as well as bending, are already known to affect the rolling process and, thus, their effect can be compensated by guiding the rolling process so that the wear and the bending of a stand become symmetrical. One of the excluded features was *the identification number of caster* and it would be extremely alarming if the two casters produced different types of slabs from the viewpoint of wedge formation. The exclusion of the feature *roll change indicator concerning strip* was somewhat unexpected, because this feature indicates if the strip was rolled right after roll change, in the middle of the roll changes, or just before roll change, and it was felt that it might affect wedge formation. This matter will be further developed in the context of parallel coordinates and k-means clustering.

The analysis of linear correlation yielded to the exclusion of two more features, i.e. *thermal crown of the sixth stand* and *thermal crown of the seventh stand*, because they included repetitious information that would later distort the results of SOM. The feature that was kept in the dataset to represent these two was *thermal crown of the fifth stand.*

The scaling of the remaining 24 features produced 24 components, which were then used to calculate the first SOM. The resulting map revealed some clusters and, on the basis of a visual inspection of the corresponding component planes, 17 components were selected for further analysis because they seemed to have an effect on wedge distribution in the clusters. Table 16 lists the seven components that were excluded, because they seemed to include repetitious information. A second SOM was calculated with the selected 17 components and, when comparing the two maps, the second map had crisper limits for clusters than

the first map. This means that the set of components that was used to calculate the second map was better adapted to differentiate the samples into separate groups than the set of components that was used to calculate the first map. The visual inspection of the resulting component planes did not reveal any need to reduce the number of components and, next, the method of parallel coordinates was used to further study the dataset.

Table 16. Components that were excluded according to visual inspection of SOMs.

Components that were excluded from further studies (number)
continuous variable crown position of the third stand, continuous variable crown position of the fifth stand, continuous variable crown position of the seventh stand, thermal crown of the first stand, thermal crown of the second stand, thermal crown of the fourth stand, thermal crown of the fifth stand (7)

When studying the parallel coordinates figure showing the distributions of samples with small wedge (with class ‘zero’, where -1.26 mils $\leq$ wedge $\leq$ 1.21 mils) in relation to samples with larger wedge (with class ‘one’), only ten components had noticeable differences. The seven components that did not have differences in their distribution and, hence, were not selected for further studies, are listed in Table 17. An interesting exclusion here was that of another indicator of roll change, which tells if any of the rolls in stand were changed before the strip in question was rolled. As it was mentioned earlier, it was suspected that the roll change would affect the wedge formation and, thus, it was tested whether the conditional probabilities could be verified for data including this component. The outcome was that the resulting probabilities could not be verified and, hence, the exclusion of the component was the correct thing to do.

Table 17. Components that were not selected according to visual inspection of parallel coordinates figure.

Components that were excluded from further studies (number)
flag of ridge detection on operator side, ridge location from edge on operator side, flag of ridge detection on driver side, ridge location from edge on driver side, continuous variable crown position of the first stand, continuous variable crown position of the second stand, roll change indicator concerning stands (7)

Consequently, the conditional probabilities were formed by using the ten selected components, which included *the average thickness of strip, the average width of strip, the difference between the measured center crown and the target center crown, the calculated wedge, the calculated crown, the height of the ridge on operator side, the height of the ridge on driver side, the continuous variable*

crown (cvc) position of the fourth stand, the cvc position of the sixth stand, and *the thermal crown of the third stand*. The calculated wedge is the difference between the edge thicknesses from the modeled strip profile. The cvc value indicates the quantity of shift between the work rolls in stand (in axis direction) and the thermal crown is the difference in radial expansion between the edge and the center of the work roll. The shifting of rolls is very important, because it helps to eliminate the problems that occur with edge profiles during the rolling of strips with similar widths. The reason for the selection of only two cvc positions and one thermal crown is that all seven cvc positions and seven thermal crowns are highly inter-correlated. A rolling expert's opinion is that the cvc correlation for multiple stands is something to be expected due to thermal crown growth appearing at the same time on work rolls and, thus, it is enough to use one or two of them to present the information that is included in all of them.

The final SOM was calculated and then clustered by using a k-means clustering algorithm. As a result, a clustering of 15 clusters was obtained, on the basis of which the conditional probabilities were formed. In order to clarify the results, two of the 15 conditional probabilities are presented below. Here, the presented values for the components are normalized and, thus, all components have for the smallest possible value zero and for the largest possible value one.

The conditional probability for the cluster number 1 is:

Probability for having larger wedge is 0.145 with conditions:

	0.25	≤	average thickness	≤ 1
and	0.22	≤	average width	≤ 0.99
and	0.02	≤	difference in center crown	≤ 0.89
and	0.08	≤	calculated wedge	≤ 0.98
and	0.29	≤	calculated crown	≤ 0.82
and	0	≤	ridge height on operator side	≤ 0.61
and	0	≤	ridge height on driver side	≤ 0.21
and	0.15	≤	continuous variable crown of 4^{th} stand	≤ 0.82
and	0.28	≤	continuous variable crown of 6^{th} stand	≤ 0.92
and	0.15	≤	thermal crown of 3^{rd} stand	≤ 0.85.

Correspondingly, the conditional probability for the cluster number 2 is:

Probability for having larger wedge is 0.011 with conditions:

	0.02	≤	average thickness	≤ 0.25
and	0.15	≤	average width	≤ 0.76
and	0.24	≤	difference in center crown	≤ 0.70
and	0.26	≤	calculated wedge	≤ 0.76
and	0.39	≤	calculated crown	≤ 0.62
and	0	≤	ridge height on operator side	≤ 0.27
and	0	≤	ridge height on driver side	≤ 0.11
and	0.10	≤	continuous variable crown of 4^{th} stand	≤ 0.55
and	0.55	≤	continuous variable crown of 6^{th} stand	≤ 1
and	0.51	≤	thermal crown of 3^{rd} stand	≤ 0.98.

When comparing these two clusters, it can be seen that the probability of having larger wedge is almost 14 times greater with conditions of cluster 1 than with conditions of cluster 2. It is also easily seen that the conditions in them differ remarkably for all ten components. For example, component 1, which is *the average thickness of strip*, has values greater than or equal to 0.25 in cluster 1 while, in cluster 2, the values are smaller than or equal to 0.25. With this component, value 0.25 corresponds to 161 mils. The calculated wedge seems to follow the true values of wedge, because it has more limited value range around the target of 0 mils when the probability for larger wedge is smaller. Actually, these values are calculated during and after rolling, and it would, thus, be disquieting if the values did not somewhat correlate with the true values. For calculated wedge, value 0.26 corresponds to –1.44 mils and value 0.76 corresponds to 1.56 mils.

After studying all the 15 conditional probabilities, the following conclusion was drawn: having high values for *thickness, width, difference in center crown, calculated wedge, calculated crown, ridge height on operator side, ridge height on driver side, cvc of 4^{th} stand*, or low values for *calculated wedge, cvc of 6^{th} stand, thermal crown of 3^{rd} stand*, produced relatively more strips with wedge smaller than -1.26 mils or larger than 1.21 mils. For example, approximately 18% of the strips in the development dataset fulfill at least one of the above-mentioned conditions and these 18% include approximately 42% of all strips with wedge smaller than -1.26 mils or larger than 1.21 mils. Thus, preventing strips to meet

these conditions would reduce by almost half the number of strips having larger wedge.

At the end of this study, the conditional probabilities were tested with new data. The evaluation dataset included the ten components present in the conditional probabilities and the data was prepared the same way as for the development dataset. Then, the samples were linked to the existing clusters and, when comparing the proportional numbers of strips with larger wedge (i.e. smaller than -1.26 mils or larger than 1.21 mils) in clusters, it was observed that the datasets had a similar distribution. Hence, the conditional probabilities were found to be reliable and these results could be used to interpret the conditions of the continuous casting process in correlation to the wedge formation.

According to this analysis, in order to obtain smaller wedge, the first eight features should have low values and the last two features should have high values. Considering the knowledge from the experts in the field, it seems rather evident that having low values for *thickness, width, variation from target center crown, calculated wedge,* and *height of the ridge* on either side would actually produce small wedge, but, on the other hand, getting small actual wedge while having, at the same time, low value for *cvc position of the fourth stand*, high value for *cvc position of the sixth stand*, and large *thermal crown of the third stand* is not as obvious a conclusion as that. This result is novel information that needs to be studied more carefully in the future. In addition, calculations confirm the already known situation that having a small calculated crown results in smaller wedge.

Even though the set of features obtained did not include casting features, except for the knowledge of which caster was used, it does not prove that casting has no influence on wedge formation. There are indeed dependencies between rolling and casting features, for example, the cvc shifting is slightly affected by the predicted caster profile. This is one of the subjects that could be further studied in this field.

6.1 Discussion

As one of the two targets of this part of the study was to prove that the derived method of conditional probabilities works also in another environment, the achieved results were very satisfying. Although Ruukki and Steel Dynamics Inc. are both steel manufacturing companies, the processes under study in this thesis were quite different from each other.

The results of this part of the study showed that the used method does not require hundreds of features to work properly, but can be used with only tens of features to start with.

Additionally, this case showed that the results of the parallel coordinates method should not be questioned based on light arguments (see Chapters 3.5 and 5.3). In fact, the lesson here would be that the conditional probabilities should be derived first after the straightforward feature selection and, then, if the results are not verified with the independent data, the part related to the parallel coordinates could be taken under more precise evaluation.

The other target of studying the Steel Dynamics data was to find the underlying conditions of the wedge formation in continuous casting hot strip rolling. Yet another successful result was obtained since, as in the case of Ruukki, the derived conditional probabilities included both facts that were already known by the experts and novel information. The outcome of the study was coherent with the expert knowledge indicating that:

- having small values for *thickness, width, variation from target center crown, calculated wedge,* and *height of the ridge* on either side would produce small actual wedge and
- having a slight calculated crown results in smaller wedge.

The gained novel information indicates that while having small value for *cvc position of the fourth stand,* large value for *cvc position of the sixth stand,* and large *thermal crown of the third stand* it would produce a small actual wedge.

Finally, the results gave two items that it would be useful to study in the future:

1. the novel information gained and
2. the dependencies between rolling and casting features as it is not proved that casting has no effect on wedge formation.

7 Results of k-NN and C4.5

This chapter presents the results derived with the methods used for the verification of the conditional probabilities method and, especially, its feature selection part. The used verification methods were k-Nearest-Neighbors and C4.5 decision tree.

7.1 k-Nearest-Neighbors

The first verification method to be presented is k-NN, which is a classification method that also includes feature selection.

7.1.1 Datasets

The datasets used with this method were selected from the most recent datasets of Ruukki (see Chapter 4.1.2) and, namely, they were thin strips with temperature related retention codes and thick strips with temperature related retention codes. One of the purposes of using other methods beside the conditional probabilities method was to test the quality of the results derived with the latter, in order to evaluate the suitability of the method in the hot strip rolling context,. Thus, k-NN was applied to subsets at all different stages of the conditional probabilities method (see Fig. 2 and Chapter 3.1). Table 18 shows the different numbers of components (i.e. normalized features) in each of the ten subsets under study.

Table 18. Used subsets and the amounts of components in them.

Stages of conditional probabilities method	Number of components	
	Subset with thin strips and temperature related retentions	Subset with thick strips and temperature related retentions
Original subset	79	77
Subset after basic statistical analysis	60	68
Subset after linear correlation analysis	60	62
Subset after self-organizing map analysis	53	47
Final subset i.e. after parallel coordinates analysis	18	16

The first column indicates the name of the method whose outcome was the number of components in the other two columns, the second column includes the number of components for the subset with thin strips and temperature related retentions, and the third column includes the number of components for the subset with thick strips and temperature related retentions. The subset with thin strips and temperature related retentions was further divided into a training set with 9493 samples and a testing set with 9494 samples. Similarly, the subset with thick strips and temperature related retentions was divided into a training and a testing set with 3002 samples each.

7.1.2 Program runs

Because the nature of k-Nearest-Neighbor calculation is very exhaustive, the method is not usually used with very large datasets. In this thesis, each of the ten subsets was examined with k-NN by using three different values for k: 3, 5, and 10. In addition, three different measures for selecting the best prediction were used:

1. The proportional error of prediction on the whole training data.
2. The proportional error of prediction on the samples with retention codes in the training data.
3. The arithmetical average of the proportional error of prediction on the samples with retention codes and the proportional error of prediction on the samples without retention codes in the training data.

These selections gave a total of 90 program runs. Fig. 21 illustrates the corresponding calculation times for thin strips with temperature related retentions by using the measure of proportional error of prediction on all samples and Fig. 22 illustrates the calculation times for thick strips with temperature related retentions by using the measure of proportional error of prediction on all samples. These calculation times also include the time used in predicting the testing data. In both figures, the horizontal axis shows the five different subsets in chronological order and the vertical axis shows the time used in calculation. Different values for k are presented with different colors and these are specified in the legends beside the figures. It should be noted that, in the figure that includes the data of thin strips (Fig. 21), time is presented in days, whereas, in the figure that includes the data of thick strips (Fig. 22), time is presented in hours. The difference here becomes from the fact that the data including thin strips has

almost 9500 samples in both the training and the testing subset and the data including thick strips has about 3000 samples in each subset. The calculations were done with Matlab in Linux environment, where it is possible to use 64-bit memory addressing as opposed to a standard personal computer and UNIX environment that use 32-bit memory addressing. Especially in the case of 9500 samples and 79 normalized components, the Linux environment was the only one where the calculations could be done.

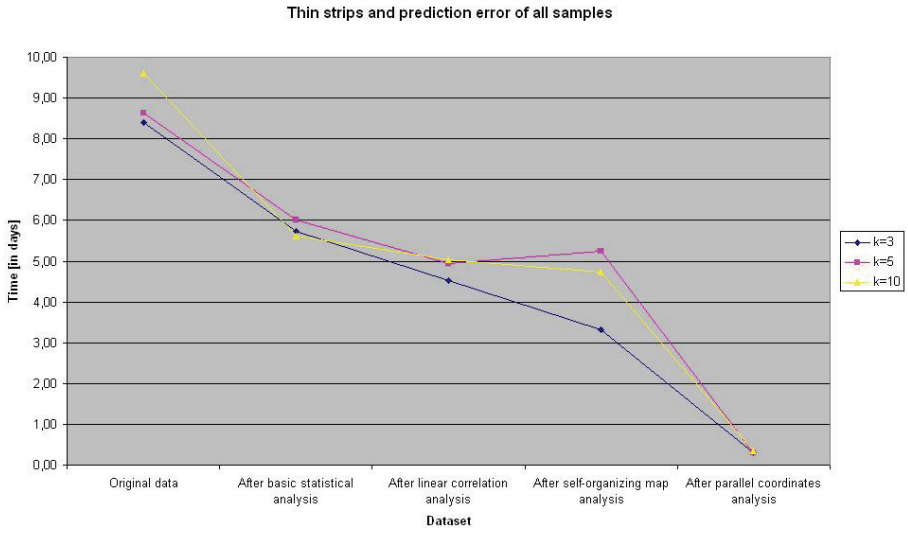

Fig. 21. Calculation times for datasets with thin strips and temperature related retentions by using the measure of proportional error of prediction on all samples.

As it can be seen from Fig. 21 and Fig. 22, the calculation time drops dramatically when the size of the feature space decreases. In the case of the thin strips, the calculations that were done with original data took more than 8 days and with the final subset (after parallel coordinates analysis) the calculation times were from 5.5 to 10.5 hours. The corresponding calculation times for data with thick strips were about 16 hours with the original data and less than a half an hour with the final subset. Thus, a very important conclusion from these calculation times is the fact that, when the feature space decreases down to 21-23% of the original size, the calculation times drop to 2-3%. The results are similar with all setting values and parameters that were used in this thesis.

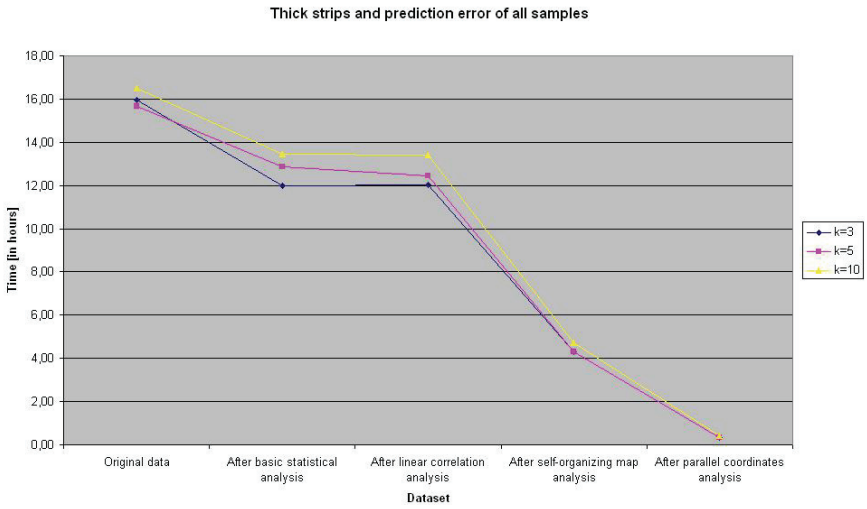

Fig. 22. Calculation times for datasets with thick strips and temperature related retentions by using the measure of proportional error of prediction on all samples.

7.1.3 Results

In this chapter, the given percentages are the results of the training datasets. This is purely due to the fact that the corresponding numbers were more easily extracted from the resulting files of the program runs. In the case of the data with thin strips and temperature related retentions, it can be noted that the prediction accuracy for the retained samples in the testing datasets was from 26 to 29% weeker than in the training datasets when the number of neighbors (=k) was equal to 3. For k=5, the prediction accuracies for the retained samples were from 14 to 17% weeker and, for k=10, from 8 to 11%. Hence, the larger the number of neighbors, the better the prediction accuracy on unseen samples followed the prediction accuracy on training data, meaning that the variance of the prediction decreases as the size of the neighborhood increases (see Chapter 3.2). Unfortunately, the prediction accuracy, i.e. bias, fall off faster than the difference between the prediction accuracy on training and testing data got smaller and, hence, the best values were always gained with number of neighbors equal to three. For the successful samples the prediction accuracy in the testing datasets differed by less than 3% when compared to the values in training datasets.

The results concerning data with thick strips and temperature retentions showed that the prediction accuracy for the retained samples in the testing datasets was from 21 to 24% less than in the training datasets when the number of neighbors was equal to 3. Correspondingly, for k=5 and k=10, the prediction accuracies for the retained samples were 16-18% and 10-12% less. Hence, also here, the variance of the prediction decreased as the size of the neighborhood increased. In cases of k being equal to 3 and 5, the bias of the prediction accuracy was growing at the same speed as the variance decreased and, thus, the best values for predicting retained samples were always similar. But for k=10, the bias worsened faster than the variance decreased and, thus, this resulted in poorer values in predictions.

In the case of the data with thin strips and temperature related retentions, the best prediction for successful samples was gained by using the measure of proportional error of prediction on the whole training data and the value 3 for k. With these settings for each of the five subsets (see Chapter 7.1.1), the best prediction was 99.9% correct by using only one component: *the minimum allowed finishing temperature*. The downside with this result was that, at the same time, the error made in predicting the retained samples was 98.1%. Nevertheless, this result shows that the component of *minimum allowed finishing temperature* is a very important factor in the case of thin strips and temperature related retentions and, as a matter of fact, it was also included in the conditional probabilities that were derived for the data in question (see Chapters 5.2.3 and 5.2.4). From future studies' and rolling expert's viewpoint, it would be interesting to look at the real temperatures measured from the rolling process and to use them to find out if the number of retentions is smaller with just a slight change in the temperature limits, such as *the minimum allowed finishing temperature*.

For the retained samples of the data with thin strips, the best prediction was achieved by using the measure of arithmetical average of proportional errors (see Chapter 7.1.2) and the value 3 for k. With these settings, the best prediction for the retained samples in the original dataset was predicted correctly at 54.1% and, at the same time, the percentage of correctly predicted successful samples was 97.9%. These percentages were achieved with the training dataset and, with the testing dataset, the corresponding prediction for retained samples was only 27.1%. The best result for the final subset, i.e. 48.9% correctly predicted retained samples, was gained by using the measure of proportional error of prediction on the retained samples and k equal to 3. Now, with the training datasets the general result was that when the successful samples were at least 98% correctly predicted

the prediction accuracy for the retained samples was about 46.6-51.3%. With the testing datasets, the corresponding prediction accuracy for the retained samples was about 72.9-78.3%. Hence, with this dataset of thin strips and temperature related retentions, it is quite impossible to employ k-NN to predict the retained samples, because tossing the coin would actually give a better result, but, on the contrary, the prediction of successful samples is quite achievable.

In the case of the data with thick strips and temperature related retentions, the best prediction of 96.6% for successful samples with training dataset was gained by using the measure of proportional error of prediction on the whole training data and the value 3 for k. This percentage was achieved with the original dataset and, at the same time, the percentage of correctly predicted retained samples was 69.2%, which could be interpreted as a sign of real prediction instead of a guess. Unfortunately, the results for the testing dataset were not as good, i.e. the prediction for successful samples was 88.2% and for retained samples only 47.4%. A more encouraging outcome was that similar percentages were achieved with all other subsets, for example, the percentages for the final subset with training dataset were 96.2% for successful samples and 69.6% for retained samples. Hence, the prediction accuracy did not change considerably while the feature space was decreased and, thus, this gives a reason to believe that the feature selection procedure does not have any impact on the level of prediction accuracy.

For the retained samples the best prediction of 82.8% was achieved by using the measure of proportional error of prediction on the retained samples and having k equal to 3, but then the percentage for the successful samples was only 41.8%. What is curious is that this result was gained by using only two components: *tolerance of allowed finishing temperature* and *deformation resistance class for pass schedule calculation = 4*. The same result was also obtained with the subsets after basic statistical analysis, after linear correlation analysis, and after self-organizing map analysis. The only difference was with the final subset, which produced the percentage of 75.7% for the retained samples and 36.8% for the successful samples with the components of *analyzed carbon content* and *tolerance of allowed coiling temperature*. This differentiation comes from the fact that the components of *deformation resistance class for pass schedule calculation* were extracted from the final subset according to the parallel coordinates analysis. Intriguingly, the component of *analyzed carbon content* stood out also when the conditional probabilities were derived (see Chapter 5.2.4). At this point, it was thought interesting to reinsert the above mentioned

components to the data (like it was done earlier with the strip production class components: see Chapter 5.2.3), but what comes to the prediction accuracies, the results were very similar to the results with the final subset. All in all, with the five subsets, when the percentage of correct prediction for the successful samples was better than 90%, the best prediction for the retained samples was between 71.0-74.7%. Thus again, the quality of the prediction remained approximately the same whether the data had 77 or 16 components.

Interestingly, especially in the case of the data with thick strips and temperature related retentions, the results of k-NN lead to search for two different sets of rules: one for predicting the retained samples and another for predicting the successful samples. Then, by using both of them either at the same time or one after the other, the final prediction accuracy might become higher. The studies on this have already been started.

During a closer study of the best results for each of the subsets with thick strips and temperature related retentions and, particularly, during the study of the components that they contained, three components stood out as being essential: *target thickness for hot strip rolling*, *strip production class = 'B'*, and *weight of slab*. The affect of *component strip production class equal to 'B'* was already discussed in Chapter 5.2 and what is most important to repeat here is that with this component, seven successfully rolled strips can be predicted correctly with 100% accuracy. In addition, as it was pointed out previously, the component of *target thickness for hot strip rolling* strongly correlates with the temperature retentions, because it is much harder to keep an even temperature within a thick strip than within a thin strip. Keeping this in mind, the results on the component of *weight of slab*, seemed at first a bit peculiar: over one third of both the lightest and the heaviest slabs, and hence of strips, were retained. The inference from this was that the extreme formats of the product are always more difficult to handle than the more often manufactured formats whose dimensions are closer to the average.

Furthermore, with the subsets of thick strips, three other components that were present in most of the best results were *deformation resistance class for pass schedule calculation equal to 2, 9*, and *12*. In order to calculate the appropriate class for a strip, the values of, for example, *calculated tensile strength of strip*, *manner of killing the steel*, and *material group* were used. Unfortunately, the calculation formulas are known only by the experts of the factory and, thus, the more detailed discussions about these exact classes could not be included in this thesis. Nevertheless, when the above-mentioned six components are compared to

the components selected in the final dataset according to the conditional probabilities method, the first three of them are the same, but the components of *deformation resistance class for pass schedule calculation* are not included. In fact, these components were excluded from the final dataset according to the parallel coordinates analysis. As it was stated in the theoretical part of the thesis (Chapter 3.1.4), the manner of using parallel coordinates that is introduced in this thesis sometimes leads to unwanted exclusion of class components. But, at least in this particular case, the results of k-NN did not indicate a significant difference between the final subsets that included the above mentioned components and the ones that did not.

When analyzing the original subset of thick strips, the best result included four components that were extracted from the data by using the basic statistical analysis (see Chapter 3.1.1). Therefore, for the sake of testing to see what would have happened if these components had been maintained in the data all the way through the feature selection procedure, they were added to the final subset along with the three components of deformation resistance class and yet another k-NN analysis was done. The results were very similar with the results gained by using the final subset that was derived with the feature selection procedure of the conditional probabilities method and, hence, these seven components had been correctly extracted from the data, because they did not include any novel or additional information in this thesis' point of view.

In conclusion, the prediction results were always similar between the different subsets when the setting parameters were the same. Thus, it shows that the feature selection procedure of the conditional probabilities method is valid and does not remove important information from the data in regards to the k-Nearest-Neighbor method. In fact, the feature selection procedure actually makes the use of k-NN more feasible and, in the case of very large datasets, when the available computing resources are not sufficient, the selection procedure is a prerequisite for using k-NN. In addition, with very large datasets and even with adequate computing resources, the time used for k-NN calculations was longer than the time used for feature selection. This includes the supposition that basic statistical analysis is required in both cases. Hence, when the k-NN analysis is demanded on large datasets, it is profitable to use the feature selection beforehand.

7.2 C4.5

The second verification method presented here is C4.5 program for creating decision trees.

7.2.1 Datasets

In order to analyze the hot strip rolling data of Ruukki with other than the developed method of conditional probabilities, the six smaller datasets, which were used for the development of the method (see Chapters 4.1.1 and 5.1), were then studied with the C4.5 method. In this case, the studies were done on each smaller dataset with two different feature quantities: 1) the quantity equal to derived conditional probabilities and 2) the quantity equal to the size of the dataset before using the parallel coordinates for feature selection. Because according to Liu and Setiono (Liu & Setiono 1996) the C4.5 works better with relevant features than with an entire dataset, it would have been desirable to study the datasets in all the different sizes that resulted from the various steps of the conditional probabilities method in order to test this statement, but the problem was that all the features were not available for the evaluation data except in the two cases mentioned above. Thus, the present analysis only tests the appropriateness of the used parallel coordinates method for feature selection in the application of the C4.5. The summary of the features included in the datasets that were used is presented in Table 19.

With the C4.5 method, the original values of features are used and, basically, the only information that the program needs in advance is to know if the used feature is either a continuous or a class feature. In the case of a class feature, the possible values have to be listed beforehand. The information that was given to the program runs about the used features is listed in Table 20. It should be noted that, in the case of continuous features, no values were given to the program beforehand, but the minimum and the maximum values are listed in the below table (Table 20) only for the purpose of analyzing the results. The given minimum and maximum values are the corresponding values of all the available data: both training and testing data.

Table 19. Summary of the datasets that were studied with C4.5 method.

Dataset	Number of features	List of features
Rolling temperature too high	18	quality of slab, charge number of slab, maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, target of thickness, target of width, upper tolerance for width, deformation resistance class for pass schedule calculation, calculated tensile strength, roughing instruction for slab, thickness of slab, length of slab, target surface quality of strip, pre-calculated length of strip, measured temperature at roughing mill
	9	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, upper tolerance for width, thickness of slab, pre-calculated length of strip, measured temperature at roughing mill
Coiling temperature close to the limits	18	(same as above 18 features in dataset of rolling temperature too high)
	6	minimum allowed finishing temperature, target of finishing temperature, thickness of slab, length of slab, pre-calculated length of strip, measured temperature at roughing mill
Too thin strip	18	(same as above 18 features in dataset of rolling temperature too high)
	8	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, thickness of slab, length of slab, measured temperature at roughing mill
Too narrow strip	18	(same as above 18 features in dataset of rolling temperature too high)
	5	minimum allowed finishing temperature, upper tolerance for width, thickness of slab, pre-calculated length of strip, measured temperature at roughing mill
Too wide strip	18	(same as above 18 features in dataset of rolling temperature too high)
	12	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, target of width, upper tolerance for width, calculated tensile strength, thickness of slab, length of slab, pre-calculated length of strip, measured temperature at roughing mill
Torn tail end	18	(same as above 18 features in dataset of rolling temperature too high)
	9	maximum allowed coiling temperature, maximum allowed finishing temperature, minimum allowed finishing temperature, target of finishing temperature, lower tolerance for thickness, calculated tensile strength, thickness of slab, pre-calculated length of strip, measured temperature at roughing mill

Table 20. Summary of the features that were used with the C4.5 method.

Feature	Type of feature	Values of feature
Quality of slab	Class	1, 2, 3, 4, 5, 6, 7, 8
Charge number of slab	Class	1, 2, 3
Maximum allowed coiling temperature	Continuous	570 – 990
Maximum allowed finishing temperature	Continuous	845 – 980
Minimum allowed finishing temperature	Continuous	700 – 910
Target of finishing temperature	Continuous	830 – 930
Lower tolerance for thickness	Continuous	0.05 – 0.46
Target of thickness	Continuous	1.41 – 15.9
Target of width	Continuous	681 – 1860
Upper tolerance for width	Class	19, 20, 25, 28.6, 30, 31.8, 35, 41.3, 48
Deformation resistance class for pass schedule calculation	Class	1, 2
Calculated tensile strength	Continuous	295 – 871
Roughing instruction for slab	Class	1, 10001, 11001, 12001, 13001, 14001, 14007, 16001, 17001, 18001, 18007, 19001, 19007
Thickness of slab	Continuous	150 – 250
Length of slab	Continuous	4200 – 11997
Target surface quality of strip	Class	0, 11, 12, 13, 14, 15, 16, 17, 18, 19
Pre-calculated length of strip	Continuous	59 – 1727
Measured temperature at roughing mill	Continuous	860 – 1177

Overall, the distribution of feature values is quite similar between the training and the testing datasets, but there was one exception related to the *pre-calculated length of strip*. For this feature, every testing dataset had much higher maximum value than the corresponding training dataset. This type of setup might lead to a situation where important information of the feature's influence on the samples is possibly not used in the creation of trees. Because the calculation of the length of strip is highly dependant on the dimensional measures of the slab, the testing datasets also had somewhat higher maximum values for the feature of *length of slab*. By looking at the samples that had higher value for *pre-calculated length of strip* in the testing dataset compared to the maximum value of the training dataset, it was found out that the size of this group was always 0.5% or less of the dataset size and most of the corresponding samples were successfully rolled strips. Hence, even though these samples could have been especially separated from the rest of the data with this method, it would not have affected the quality of the results in a significant way. Thus, the difference between the maximum value of

pre-calculated length of strip in the training and testing datasets was considered not to be significant.

7.2.2 Program runs

When using the C4.5 program, four different runs were done for each dataset. All runs complied with the following four options:

- Testing (or evaluation) with unknown samples.
- Tree has to have at least two leaves that have at least two samples each.
- The gain ratio criterion was used for evaluating the tests.
- Confidence level of 25% was used.

The evaluation with unknown samples is utilized in this study, because the data available is sufficient for it and because the utilization of unknown samples for testing yields more reliable results. This option is just induced in the program for cases where enough data is not available for testing with unknown samples. As mentioned in Chapter 7.2.1, there was no data available for evaluation purposes with larger feature sizes. This problem could have been ignored by not choosing the option of unknown samples, but then the results would have most probably been misleading.

The option of having at least two leafs that have at least two samples each, was selected to make sure that the two possible extremes for the shape of a tree are avoided. In other words, this option ensures that the resulting tree is not only a single leaf or built in such a way that a separate leaf represents each sample.

The gain ratio criterion was chosen over gain criterion, because it should work better according to the developer of the program (see Chapter 3.3). The fourth option in the above list, namely the value of the confidence level, has an effect on the decision tree pruning. This value is used in the estimation of the error rate of unseen samples and small values cause heavier pruning than large values. Quinlan suggests that one should decrease the value if the actual error rate of pruned trees on test samples is much higher than the estimated error rate (indicative of underpruning). (Quinlan 1993) In this thesis, the confidence level of 25% was selected, because it was the default value for the program and it seemed to work reasonably well.

In addition to the above mentioned options, three other options were used in turns: grouping the feature values, soft thresholds, and windowing. These options are explained in more detail in Chapter 3.3. In conclusion, four different runs

were done for each of the twelve datasets and the selected options for each of the four runs are shown in Table 21.

Table 21. Summary of the options that were used while running the C4.5 program.

Run	Basic options	Additional option
First	Evaluation with unknown samples.	None.
	At least two leaves that have at least two samples each.	
	Gain ratio criterion used for evaluating the tests.	
	Confidence level of 25%.	
Second	Evaluation with unknown samples.	Grouping of the discrete values.
	At least two leaves that have at least two samples each.	
	Gain ratio criterion used for evaluating the tests.	
	Confidence level of 25%.	
Third	Evaluation with unknown samples.	Soft thresholds.
	At least two leaves that have at least two samples each.	
	Gain ratio criterion used for evaluating the tests.	
	Confidence level of 25%.	
Fourth	Evaluation with unknown samples.	Windowing with 20 windows.
	At least two leaves that have at least two samples each.	
	Gain ratio criterion used for evaluating the tests.	
	Confidence level of 25%.	

7.2.3 Results

At first, the decision trees were formed for six datasets with 18 features. All in all, the C4.5 method did not prove to be a very good method for analyzing the hot strip rolling data, this conclusion being based on the fact that the resulting trees did not succeed to classify the samples into retained and successful strips. As an example of the resulting trees, Table 22 and Table 23, respectively, show the best and the worst results with these datasets. Here, the quality of the classification is measured by the proportion of the correctly classified retained strips. In the tables below, the upper part indicates the sizes and the error rates of the trees for both training and testing datasets and in the case of both the unpruned and the pruned tree. The upper part also shows the estimated error rate for the tree after pruning. The lower part shows how both the successful strips and the retained strips of the testing dataset were classified by the pruned tree.

Table 22. The best result gained with 18 features by using C4.5: dataset includes the retentions of coiling temperature close to the limits.

Dataset	Tree before pruning		Tree after pruning		
	Number of nodes	Incorrectly classified samples (%)	Number of nodes	Incorrectly classified samples (%)	Estimated error rate %
Training data	2798	1568 (6.7%)	954	2051 (8.8%)	11.4%
Testing data	2798	4863 (17.9%)	954	3999 (14.7%)	11.4%

True class		Classification by pruned tree	
		(a)	(b)
Successful strip (24499 samples)	(a)	22424	2075
Retained strip (2633 samples)	(b)	1924	709

Table 22 shows the result of the fourth run (see Chapter 7.2.2) for the data with 18 features and retentions of coiling temperature close to the limits. The results of the classification showed that 709 retained strips, i.e. 26.9% of the total of 2633 retained strips, were classified correctly. This was the best result gained with datasets that had 18 features and it was not very encouraging. Sometimes, when making predictions on the outcome of the rolling process, even the knowledge of finding every fourth possible retention might be a very good result, but there are other aspects to be considered too. For example, the tree in question had two important downsides, the first of which was that the size of the tree after pruning was still enormous (954 nodes). As the purpose of the study was not just to find some type of classifier for the data, but rather obtain knowledge of the underlying connections between different features and their effect on retention occurrence, the large size of the tree made it quite impossible to analyze all the rules that it contained, but, then again, just by looking at certain parts of the tree it was possible to find some kind of insight into the data. For example, by selecting 25 of the derived rules, the estimated error rate was only 3.1% for classifying 60.8% of the strips from the dataset, but, then again, most of these strips would have been classified as successfully rolled and it would have been desirable to also gain information from the retained strips. The second downside of the tree in question was that it classified incorrectly 2075 successful strips. On the other hand, having 8.5% (=2075/24499) of the successful strips classified as retained might actually be acceptable, but only if all (or most) of the retained strips were classified correctly. The analogy behind this is that, supposedly, there is a classifier working on-line in the rolling process that predicts if the current slab to

be rolled is going to be retained or not. Now, if the retained strips could be classified quite reliably, it might be worthwhile to spend some more time during the rolling process to control the manufacturing of these slabs into strips and, hence, maybe reduce the number of retentions. The opposite would then be to have a classifier that cannot detect the possible retentions and, hence, all slabs should be more carefully rolled. Thus, because the retained strips were rather far from perfectly classified, the result was not useful in the steel manufacturer's point of view.

Table 23 shows the result of the first run for the data with 18 features and retentions of too thin strips.

Table 23. The worst result gained with 18 features by using C4.5: dataset includes the retentions of too thin strips.

Dataset	Tree before pruning		Tree after pruning		
	Number of nodes	Incorrectly classified samples (%)	Number of nodes	Incorrectly classified samples (%)	Estimated error rate %
Training data	535	186 (0.9%)	5	271 (1.4%)	1.4%
Testing data	535	949 (3.7%)	5	828 (3.3%)	1.4%

True class		Classification by pruned tree	
		(a)	(b)
Successful strip (24499 samples)	(a)	24499	0
Retained strip (828 samples)	(b)	828	0

The results of the classification revealed that all the strips were classified as successful. Even though the pruned tree had 5 nodes, it could have been replaced with a tree that had only one leaf, which would then classify the strips as successful. Because all the other runs with this dataset had similar results, it suggested that there were not enough samples of retained strips in proportion to the number of successful strips in order for this type of retention to be classified with the C4.5 decision tree. The error rate of 3.3%, which was the percentage of retentions in the testing dataset in question, was just so small that it was very hard to find a classification that would produce a better error rate.

Considering two of the datasets, namely the ones with retentions of rolling temperature too high and coiling temperature close to the limits, a slightly more encouraging outcome was that it was possible to separate a few individual rules from the resulting decision trees. In the case of the dataset with retentions of rolling temperature too high, all samples that had the value of *maximum allowed*

finishing temperature within its uppermost third were successfully rolled. Then again, this could have been noticed more easily, for example, from the simple histogram of this feature. It is also logical that, if the value of *maximum allowed finishing temperature* is set high enough, there will be no strips with higher temperature and, thus, there will not be any retentions for too high rolling (i.e. finishing) temperature. In the case of the dataset with retentions of coiling temperature close to the limits, the retentions were more likely if the *maximum allowed coiling temperature* had a value within its lowermost fifth, except if the value was very close to feature's minimum and, at the same time, the *minimum allowed finishing temperature* had a value that was smaller than the feature's average value. Again, the first part of the rule could have been seen, for example, from the histogram and it was a very logical result as the lower value of *maximum allowed coiling temperature* will most likely to be exceeded and, thus, retentions of this type are more easily produced. On the other hand, the latter part of the rule was somewhat more complex, but also very natural in this context, because, usually, the lower the *minimum allowed finishing temperature* the lower the *maximum allowed finishing temperature* and, thus, if the temperature of the rolled strip is between these two limits, then, the *lower maximum allowed coiling temperature* is more probably surpassed.

An overall result for the dataset with eighteen features was that it was possible to find a set of rules that would classify approximately 60% of the strips correctly with a small estimated error rate, but, usually, these rules produced the class label of successfully rolled strips. Thus, the final results of this analysis were deficient in the sense that knowledge about the retained strips would have been appreciated too.

Next, the decision trees were formed for six datasets that had from five to twelve features according to the final results of feature selection with the conditional probabilities method (see Table 19). Table 24 shows the result of the first run for the data with six features and retentions of coiling temperature close to the limits. The results of the classification showed that 605 retained strips, i.e. 22.8% of 2651 retained strips, were classified correctly. Even though the number of nodes in the pruned tree (387) was now only about 40% percent when compared to the corresponding tree with eighteen features (954 nodes), the actual size was still too large for extracting specific rules for analysis purposes. What is more, the percentage of incorrectly classified successfully rolled strips (13.6% = 3380/24781) was now 1.6 times the percentage obtained with the dataset including eighteen features (8.5%).

Table 24. The best result gained with smaller amount of features by using C4.5: dataset includes six features and the retentions of coiling temperature close to the limits.

Dataset	Tree before pruning		Tree after pruning		
	Number of nodes	Incorrectly classified samples (%)	Number of nodes	Incorrectly classified samples (%)	Estimated error rate %
Training data	791	2579 (11.0%)	387	2752 (11.8%)	13.5%
Testing data	791	5790 (21.1%)	387	5426 (19.8%)	13.5%

True class		Classification by pruned tree	
		(a)	(b)
Successful strip (24499 samples)	(a)	21401	3380
Retained strip (2651 samples)	(b)	2046	605

In general, when comparing these results with the classifications of the datasets with eighteen features, it was found out that the decision trees were smaller and more comprehensible, but, at the same time, the performance of the trees was, at best, the same and, in most cases, somewhat worse. Hence, as it was already stated in Chapter 3.1.4 and Chapter 5.2.3, the method of using parallel coordinates for feature selection might not work well in some cases and, when using C4.5 program with these datasets, this seemed to be true because the results were worse with the data obtained after the use of parallel coordinates in contrast to the data obtained before using it.

7.3 Discussion

Here, two methods, namely, k-Nearest-Neighbor and C4.5, were used to verify the quality of the derived feature selection method. In case of k-NN, the results clearly showed that when a constant setup of parameters was given for the k-NN, the level of the outcome stayed the same even though the input feature space was decreased step by step, according to the derived feature selection procedure. This proves that the used feature selection is capable to extract features that do not affect the outcome of the classifier. In the case of large feature space (about 80 features) and quite large sample space (about 3000 samples in both training and testing data), when the feature space was decreased into 21% of the original size, the calculation times of the classification dropped enormously, i.e. by 98% of the original time spent (from about 16 hours to about 20 minutes). The situation was

even better when the sample space was about 9500 samples in both training and testing data, because then the calculation times dropped from the original time of 8 to 10 days down to 5 to 10 hours. This gain in time is surely useful, because the fine-tuning of the classification then takes less time than the process of feature selection and, hence, the overall time for analysis is reduced.

The results of this part of the study showed that with C4.5 it was possible to extract specific rules for classification. The downside was that with the used datasets, the results included only already known facts about the processes and no novel information. On the other hand, the results obtained with the k-NN, indicating that it was not possible to extract specific rules for classification, clearly showed that this method is not capable of handling this type of data that includes very complex interactions between different features.

It could have been interesting to test C4.5 after each stage in the feature selection process, starting with the original data, but this was not possible since the evaluation data was only available for features that were left in the data after using SOM for feature selection. Now, the reason why C4.5 worked better with the data before parallel coordinates was used might lie in the way a decision tree is built. It seems that the program handles better features that have restricted values as opposed to features that have continuous values and, when the parallel coordinates was used, most of the extracted features were actually class features. This characteristic of a classification method is quite unwanted in the viewpoint of hot strip rolling, because most of the measures taken from the process are actually continuous variables.

Nevertheless, in the future, it might be an attractive line of research related to C4.5 to use groups of features in making the decision nodes instead of looking at one at a time. Maybe this would help to diminish the influence of just one feature and render the results more eligible. In case of k-NN, this was taken into account by having the possibility to extract a feature from the already selected feature space (see Chapter 3.2). But how then to select the suitable groups of features for C4.5? This definitely needs to be investigated in more detail.

8 Conclusions

This chapter presents the discussion and the conclusions on the study as a whole.

8.1 Discussion

At present, various processes in industry are becoming more automated and complex. This means that process control is becoming more difficult as well. As a consequence, knowledge discovery methods are becoming ever more vital, because they can provide a way to comprehend the underlying relationships between the process parameters. The purpose of this thesis was to develop a novel method of exploring large datasets, which was, at the same time, applied to two different cases of hot strip rolling. At first, the method was developed by studying the occurrence and the absence of retentions during a hot strip rolling process. In the past, the main barrier to this type of research has been the lack of knowledge of how to utilize the enormous databases resulting from measurements taken from the rolling process. The reason for seeking the solution is obvious: if a product has to be rechecked after rolling, the productivity of the factory is decreased while the costs are increased in many ways.

As a result of the first part of the study, a novel method was developed permitting the analysis of large datasets by using also the knowledge of experts on the problem. With regards to the retention dataset, this method helped to reduce the size of the original dataset to 14%, which, in this case, meant that tens of features still remained to be studied. In many cases, such a drastic reduction might not be needed, but even a small cut could make a difference in the result. Later on, by using the combination of SOMs, parallel coordinates and k-means clustering, the developed method helped to determine the conditions of the appearance and the disappearance of four common types of defects. Without the use of SOMs and parallel coordinates together, the resulting rules would have been too complicated. At this point, the results helped the rolling experts to get new insight into hot strip rolling and, as a consequence, a new dataset was gathered in order to obtain even more specific information.

The more recent dataset including the retention information was gathered by using the knowledge gained from the first part of the study. This time, only a little over fifty features were selected for the preliminary dataset based on the earlier results and the discussions with the rolling experts, while, in the first case, the preliminary dataset included almost 240 features. Also, the subsets were now

formed in a better-considered way, because the knowledge from the previous part of the study helped to take into account the effects of a very dominant feature that was revealed by the previous results and, this way, it was possible to obtain even better results. What is meant by better results is that, when the data was studied without taking into account the effect of the dominant feature, it was not possible to verify a part of the results, whereas, when the data was preprocessed by using the knowledge of the dominant feature, it was possible to verify results for each of the subsets that were studied. In this case, the rolling experts' opinion was essential in deciding how to handle the dominant feature and, thus hopefully, this example will encourage the researches and the experts to join their efforts. In the future, the clusterings that were obtained can be used when a new sample is collected, by linking the sample to the best matching cluster, which then gives a probability for having a certain type of retention. This information can then be used to find the process parameters that will most likely produce a successful strip.

From the theoretical point of view, the goal of the continuous casting study was twofold: to show that the method for deriving conditional probabilities would work with smaller datasets and, that it would not be tied into one type of data. The results of this thesis support these assumptions and, thus, both of the goals were achieved. Since the used method is not tied into a specific type of data, it can, in the future, be used to study other datasets that include vast numbers of features and samples. Another aspect is that, although the method is able to handle hundreds of features together, it does not require such a volume to work properly and it can, thus, be used in cases with fewer features.

From the application's point of view, the purpose of this thesis was to discover which features affect the formation of the wedge. The achieved information broadens the expert knowledge of the process and will, thus, help to make more conscious use of the rolling. As a result, the method used here revealed the rolling conditions for the occurrence of wedge. The analysis shows that the most important features affecting wedge formation are *the average thickness of strip, the average width of strip, the difference between measured center crown and the target center crown, the calculated wedge, the calculated crown, the height of the ridge on operator side, the height of the ridge on driver side, the cvc position of the fourth stand, the cvc position of the sixth stand*, and *the thermal crown of the third stand*. When compared to the earlier studies (Biggs *et al.* 2000, Loney *et al.* 2002, Mücke *et al.* 2002, Shiraishi *et al.* 1991,

Tarnopolskaya *et al.* 2002), this set of ten features is quite different and this information alone gives novel insight into the process under study.

An interesting additional result is that the selection of the caster is not significant in wedge formation's point of view. This result supports the assumption that the casters perform similarly, but, on the other hand, it does not exclude the idea that the continuous casting process could produce wedge on slab. If the slab profile has a wedge, it is difficult to get rid of it in the final product. For this thesis, the wedge of the slab was not measured, which makes this a possibly interesting aspect to study further.

Sometimes, when one is handling very large datasets, the available memory and computing capacity prove to be too small. At such a time, it is convenient to have a method of processing the data in such way that the volume can be decreased into feasible size without loosing any important information. There are two ways to do this: either use a smaller sample size or decrease the feature space. In this thesis, a novel procedure of feature selection was presented and its quality was then verified with the k-Nearest-Neighbor method. Over ninety program runs were made with Ruukki data and the calculation times were recorded. From these measured times, it was concluded that, when the feature space decreased down to 21-23% of the original size, the corresponding calculation times dropped to 2-3% at the same time. Of course, this advantage in computing has no importance if the results of the analysis are not on the same level. Hence, it was extremely gratifying to discover that the prediction results were always similar between the different subsets and, thus, the feature selection procedure of the conditional probabilities method was found to be valid in the sense that it does not remove important information from the data. Consequently, the presented feature selection procedure enables the utilization of more versatile methods when the capacity of computing does not have such an impact on the decision on which method to use for the study.

The studies of k-Nearest-Neighbors and C4.5 prove that the case of hot strip rolling is very complex and tough to be modeled. In particular, when using k-NN and C4.5 that are the type of methods that analyze a single feature at a time, it is difficult to grasp the more complicated inter-correlations in the data. In the case of the k-NN, the discovery of the global minimum (or maximum) instead of the local one could be achieved with an exhaustive search through all the possible combinations of features, but this is simply not feasible, because the method's computational load is very high. In the case of the C4.5, the results were somewhat deficient in a sense that knowledge about the inter-correlations

between features in the case of the retained strips could not be derived from the resulting trees, but, on the other hand, it was possible to extract rules that classified about 60% of the strips correctly, with a small estimated error rate. These rules contained, for example, logical connections between different temperature features and, thus, they were considered to be valid. Nevetheless, especially in the case of the hot strip rolling data with thick strips and temperature related retentions, the results of k-NN point to a direction where two sets of rules could be used to analyze the data: one for predicting the retained strips and another for predicting the successful strips. This idea is definitely worthwhile to study in the future.

8.2 Summary

The first chapter of this thesis introduced the knowledge discovery process in general, including some methods used in the process, and presented a few interesting targets of application. The real contribution of this thesis was presented starting from the Chapter 2, which, mainly, outlined how the process of knowledge discovering was assembled to be utilized in the study. The presented process has two main divergences from the earlier processes: first, it interlocks the different steps of the process instead of keeping them totally separate and, second, it emphasizes the value of taking advantage of expert knowledge of the problem.

In Chapter 3, the theories behind the used methods were depicted more closely. This chapter also outlined the strong and the weaker points of the methods. Fortunately, the strong points were overwhelming and, hence, this gave confidence in the usability of the developed process. The main contributions of the novel method presented in this thesis, in terms of the used methods, are that expert knowledge can be utilized in the process, that the datasets under study can be enormous in both feature and sample dimensions, that the used methods are powerful visualization tools (and, thus, utilize the inborn human ability to find patterns), and that the form of the results is very unrestrictive and easily interpreted. Another positive aspect is that the instructions also include some specific limits to be used as a starting point for feature selections with large datasets.

Chapter 4 presented the targets of application. The main distinction from the earlier studies was that the processes are investigated as an entity instead of in smaller segments. In addition, this chapter outlined the differences in the target

applications so that the requirement of testing the developed novel method in a different environment could be fulfilled.

In regards to the results that were obtained with the developed method in two different environments, they were presented in the fifth and sixth chapters. The results show that the derived method is fulfilling its target as a knowledge discovery process since it was able to find novel and useful information from all the cases under study. What is more, the content of Chapter 6 proved the following two assumptions on the derived method:

1. the method works with smaller datasets that do not include hundreds of features and
2. the method is not tied into one specific type of data.

Finally, Chapter 7 described the results that were derived with methods of k-Nearest-Neighbors and C4.5. The main purpose of using these methods was to provide verification for the quality of the derived feature selection procedure and, as the results showed, the quality was verified. Hence, the overall conclusion was that the derived method is valid and that it should be used because it renders the interpretation of the complex inter-correlations within data more feasible.

References

Anan G, Nakajima S, Miyahara M, Nanba S, Umemoto M, Hiramatsu A, Moriya A & Watanabe T (1992) A Model for Recovery and Recrystallization of Hot Deformed Austenite Considering Structural Heterogeneity. ISIJ International, 32(3): 261-266.

Andrienko G & Andrienko N (2004) Parallel Coordinates for Exploring Properties of Subsets. In: Proceedings of Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'04). IEEE Computer Society, pp. 93-104.

Apte C & Hong SJ (1996) Predicting Equity Returns from Securities Data with Minimal Rule Generation. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P & Uthurusamy R (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, California, USA, pp. 541-560.

Bendat J & Piersol A (1971) Random Data: Analysis and Measurement Procedures. John Wiley & Sons Inc., USA.

Bhadeshia HKDH (1999) Neural Networks in Materials Science. ISIJ International 39(10): 966-979.

Biggs DL, Hardy SJ & Brown KJ (2000) Influence of Process Variables on Development of Camber during Hot Rolling of Strip Steel. Ironmaking and Steelmaking 27(1): 55-62.

Bloch G, Sirou F, Eustache V & Fatrez P (1997) Neural Intelligent Control for a Steel Plant. IEEE Transactions on Neural Networks 8(4): 910-918.

Blum AL & Langley P (1997) Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence 97(1-2): 245-271.

Campos AM, García DF, De Abajo N & González JA (2004) Real-Time Rule-Based Control of the Thermal Crown of Work Rolls Installed in Hot Strip Mills. IEEE Interactions on Industry Applications 40(2): 642-649.

Caruana R & Freitag D (1994) Greedy Attribute Selection. In: Proceedings of the Eleventh International Conference on Machine Learning. Morgan Kaufmann, New Brunswick, NJ, USA, pp. 28-36.

Chang J, Remmen H, Ward W, Regnier F, Richardson A & Cornell J (2004) Processing of Data Generated by 2-Dimensional Gel Electrophoresis for Statistical Analysis: Missing Data, Normalization, and Statistics. Journal of Proteome Research 3(6): 1210-1218.

Cheeseman P & Stutz J (1996) Bayesian Classification (AutoClass): Theory and Results. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P & Uthurusamy R (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, California, USA, pp. 153-180.

Chen JX & Wang S (2001) Data Visualization: Parallel Coordinates and Dimension Reduction. Computing in Science and Engineering 3(5): 110-113.

Chou S-, Lin S- & Yeh C- (1999) Cluster Identification with Parallel Coordinates. Pattern Recognition Letters 20(6): 565-572.

Cser L, Korhonen AS, Gulyás J, Mäntylä P, Simula O, Reiss G & Ruha P (1999) Data Mining and State Monitoring in Hot Rolling. In: Meech JA & et al. (eds.) Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials (IPMM'99). IEEE , Honolulu, Hawaii, pp. 529-536.

Cser L, Simula O, Ruha P & Arvai L (2001) Quality Prediction Based on Knowledge Extraction in Hot Rolling. In: Proceedings of 3rd International Conference on Intelligent Processing and Manufacturing of Materials (IPMM 2001). IEEE, Vancouver, Canada.

Dash M & Liu H (1997) Feature Selection for Classification. Intelligent Data Analysis 1(3): 131-156.

Davies DL & Bouldin DW (1979) A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1(2): 224-227.

Dumortier C & Lehert P (1999) Statistical Modelling of Mechanical Tensile Properties of Steels by Using Neural Networks and Multivariate Data Analysis. ISIJ International 39(10): 980-985.

Famili A, Shen W-, Weber R & Simoudis E (1997) Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis 1(1): 3-23.

Fayyad UM, Djorgovski SG & Weir N (1996) Automating the Analysis and Cataloging of Sky Surveys. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P & Uthurusamy R (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, California, USA, pp. 471-493.

Femminella OP, Starink MJ, Brown M, Sinclair I, Harris CJ & Reed PAS (1999) Data Pre-Processing/Model Initialisation in Neurofuzzy Modelling of Structure-Property Relationships in Al-Zn-Mg-Cu Alloys. ISIJ International 39(10): 1027-1037.

Frayman Y, Rolfe BF & Webb GI (2002) Solving Regression Problems Using Competitive Ensemble Models. In: McKay B & Slaney J (eds.) Lecture Notes in Computer Science. Canberra, Australia. Springer Berlin, Heidelberg, pp. 511-522.

Fua Y-, Ward MO & Rundensteiner EA (1999) Hierarchical Parallel Coordinates for Exploration of Large Datasets. In: Proceedings of the Conference on Visualization '99: Celebrating Ten Years. San Francisco, California, USA. IEEE Computer Society Press, Los Alamitos, California, USA, pp. 43-50.

Ginzburg V (1993) High-Quality Steel Rolling: Theory and Practice. Marcel Dekker, Inc., New York, New York, USA.

Glick M & Hieftje GM (1991) Classification of Alloys with an Artificial Neural Network and Multivariate Calibration of Glow-Discharge Emission Spectra. Applied Spectroscopy 45(10): 1706-1716.

Gorni AA (1997) The Application of Neural Networks in the Modeling of Plate Rolling Processes. JOM-e the electronic supplement to JOM 4. 20.09.2005 Cited 06.10.1999 from: http://www.tms.org/pubs/journals/JOM/9704/Gorni/Gorni-9704.html, referenced 11.7.2006.

Goser KF (1997) Self-Organizing Maps for Intelligent Process Control. In: Proceedings of Workshop on Self-Organizing Maps (WSOM'97). Espoo, Finland, pp. 75-79.

Grošelj N, Zupan J, Reich S, Dawidowski L, Gomez D & Magallanes J (2004) 2D Mapping by Kohonen Networks of the Air Quality Data from a Large City. Journal of Chemical Information and Computer Sciences 44(2): 339-346.

Gyenesei A (2004) Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes. Doctoral Thesis. University of Turku, Department of Information Technology, Turku, Finland, 60 p.

Hair JF, Anderson RE, Tatham RL & Black WC (1995) Multivariate Data Analysis: with Readings. 4th edition. Prentice-Hall, New Jersey, USA.

Hall LO & Berthold MR (2000) Fuzzy Parallel Coordinates. In: Proceedings of 19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS). IEEE, Atlanta, Georgia, USA, pp. 74-78.

Hall MA (2000) Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: Langley P (ed.) Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000). Stanford University, USA. Morgan Kaufmann Publishers, San Francisco, California, USA, pp. 359-366.

Hall MA & Holmes G (2003) Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering 15(6): 1437-1447.

Hastie T, Tibshirani R & Friedman J (2001) The Elements of Statistical Learning. Springer, New York, New York, USA.

Huang Z, Pei M, Goodman E, Huang Y & Li G (2003) Genetic Algorithm Optimized Feature Transformation - A Comparison with Different Classifiers. In: Proceedings of International Conference on Genetic and Evolutionary Computation (GECCO 2003). Chicago, IL, USA. Springer-Berlin, Heidelberg, pp. 2121-2133.

Inselberg A (2002) Visualization and Data Mining of High Dimensional Data. Chemometrics and Intelligent Laboratory Systems 60: 147-159.

Inselberg A (1998) Visual Data Mining with Parallel Coordinates. Journal of Computational Statistics 13(1): 47-63.

Inselberg A & Avidan T (1999) the Automated Multidimensional Detective. In: Proceedings of the 1999 IEEE Symposium on Information Visualization (InfoVis'99). IEEE Computer Society, Washington, DC, USA, pp. 112-119.

Junno S (1989) Optimization of Steel Mill Production. Acta Universitatis Ouluensis, Series C Technica, 49: 72 p.

Kaski S (1997) Data Exploration Using Self-Organizing Maps. 56 p.

Keim DA (2002) Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics 8(1): 1-8.

Keim DA (1995) Visual Support for Query Specification and Data Mining. : Verlag Shaker, Aachen, Germany.

Kohavi R & John GH (1997) Wrappers for Feature Subset Selection. Artificial Intelligence 97(1-2): 273-324.

Kohonen T (1995) Self-Organizing Maps. Springer-Verlag, Heidelberg.

Laaksonen J & Oja E (1996) Classification with Learning k-Nearest Neighbors. In: IEEE International Conference on Neural Networks. Washington, DC, USA, pp. 1480-1483.

Laine S (2003) Using visualization, variable selection and feature extraction to learn from industrial data. Doctoral Thesis. Helsinki University of Technology, Helsinki University of Technology Publications in Computer and Information Science, Espoo, Finland, 56 p.

Lin T-, Li H- & Tsai K- (2004) Implementing the Fisher's Discriminant Ratio in a *k*-Means Clustering Algorithm for Feature Selection and Data Set Trimming. Journal of Chemical Information and Computer Sciences 44(1): 76-87.

Liu H & Setiono R (1996) A Probabilistic Approach to Feature Selection - A Filter Solution. In: Proceedings of 13th International Conference on Machine Learning (ICML'96). Bari, Italy, pp. 319-327.

Loney DW, Neuschutz E, Santifilippo F, Robinson JJ & Nilsson A (2002) Rolling Flat Products - Optimisation of the Use of a Strip Geometry Control System. European Commission, Luxembourg.

Louhenkilpi S (ed.) (1990) Continuous Casting of Steel: 1984-1989 = Teräksen jatkuvavalu: 1984-1989. TEKES, Helsinki, Finland.

Lumijärvi J (2003) Läheisyysfunktioista sekamuotoisen datan luokittelussa. Pro Gradu, University of Tampere, 75 p.

Mannila H (1997) Methods and Problems in Data Mining (a Tutorial). In: Afrati FN & Kolatis PG (eds.) Proceedings of International Conference on Database Theory (ICDT'97). Delphi, Greece. Springer, Heidelberg, pp. 41-55.

Masui T, Kaseda Y & Isaka K (2000) Basic Examination on Strip Wandering in Processing Plants. ISIJ International 40(10): 1019-1023.

Matheus CJ, Chan PK & Piatetsky-Shapiro G (1993) Systems for Knowledge Discovery in Databases. IEEE Transactions on Knowledge and Data Engineering 5(6): 903-913.

Matheus CJ, Piatetsky-Shapiro G & McNeill D (1996) Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P & Uthurusamy R (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, California, USA, pp. 495-515.

McCombie G, Staab D, Stoeckli M & Knochenmuss R (2005) Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis. Analytical Chemistry 77(19): 6118-6124.

Michie D, Spiegelhalter DJ & Taylor CC (eds.) (1994) Machine Learning, Neural and Statistical Classification. Prentice Hall.

Mitchell T (1997) Machine Learning. The McGraw-Hill Companies, Inc., USA.

Mücke G, Karhausen KF & Pütz P- (2002) Methods of Describing and Assessing Shape Deviations in Strips. MPT International 3: 58-65.

Panigrahi BK (2001) Processing of Low Carbon Steel Plate and Hot Strip – An Overview. Bulletin of Materials Science, Indian Academy of Sciences, 24(4): 361-371.

Pican N, Bresson P, Alexandre F, Haton J- & Couriot E (1993) A Perceptron with Optimized Backpropagation Learning Algorithm to Preset a Temper Mill Machine: Neuroskin. In: Proceedings of 6th International Conference on Neural Networks and Their Industrial and Cognitive Applications (Neuro-Nimes '93). Nimes, France, pp. 17-24.

Portmann NF, Lindhoff D, Sorgel G & Gramckow O (1995) Application of Neural Networks in Rolling Mill Automation. Iron and Steel Engineer, pp. 33-36.

Pudil P & Novovičová J (1998) Novel Methods for Subset Selection with Respect to Problem Knowledge. IEEE Intelligent Systems 13(2): 66-74.

Pyle D (1999) Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc., San Francisco, California, USA.

Quinlan JR (1996) Improved Use of Continuous Attributes in C4.5. Journal of Artificial Intelligence Research 4: 77-90.

Quinlan JR (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, California, USA.

R-Roda I, Comas J, Poch M, Sanchez-Marre M & Cortes U (2001) Automatic Knowledge Acquisition from Complex Processes for the Development of Knowledge-Based Systems. Industrial & Engineering Chemistry Research 40(15): 3353-3360.

Rygielski C, Wang J- & Yen DC (2002) Data Mining Techniques for Customer Relationship Management. Technology In Society 24: 483-502.

Saxén H, Lassus L, Seppänen M & Karjalahti T (2000) Pattern Recognition and Classification of Blast Furnace Wall Temperatures. Ironmaking and Steelmaking 27(3): 207-211.

Sbarbaro-Hofer D, Neumerkel D & Hunt K (1993) Neural Control of a Steel Rolling Mill. IEEE Control Systems 13(3): 69-75.

Setiono R & Liu H (1997) Neural-Network Feature Selector. IEEE Transactions on Neural Networks 8(3): 654-662.

Shahin MA & Symons SJ (2001) A Machine Vision System for Grading Lentils. Canadian Biosystems Engineering 43: 7.7-7.14.

Shiraishi T, Ibata H, Mizuta A, Nomura S, Yoneda E & Hirata K (1991) Relation between Camber and Wedge in Flat Rolling under Restrictions of Lateral Movement. ISIJ International 31(6): 583-587.

Siirtola H (2003) Combining Parallel Coordinates with the Reorderable Matrix. In: Proceedings of International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'03). IEEE Computer Society, 63-74.

Singh SB, Bhadeshia HKDH, Mackay DJC, Carey H & Martin I (1998) Neural Network Analysis of Steel Plate Processing. Ironmaking and Steelmaking 25(5): 355-365.

Smyth P, Burl MC, Fayyad UM & Perona P (1996) Modeling Subjective Uncertainty in Image Annotation. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P & Uthurusamy R (eds.) Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, California, USA, pp. 517-539.

Tarnopolskaya T, de Hoog FR, Gates DJ, Dixon A & Yuen WYD (2002) Analysis of Strip Track-Off during Flat Rolling. In: Proceedings of 44th Mechanical Working and Steel Processing Conference (MWSP). Orlando, Florida, USA, pp. 237-246.

Tetko IV, Solov'ev AV, Antonov AV, Yao X, Doucet JP, Fan B, Hoonakker F, Fourches D, Jost P, Lachiche N & Varnek A (2006) Benchmarking of Linear and Nonlinear Approaches for Quantitative Structure-Property Relationship Studies of Metal Complexation with Ionophores. Journal of Chemical Information and Modeling 46(2): 808-819.

Tryba V & Goser K (1991) Self-Organizing Feature Maps for Process Control in Chemistry. In: Kohonen T, Mäkisara K, Simula O & Kangas J (eds.) Artificial Neural Networks. Elsevier Science Publishers B.V., Amsterdam, Netherlands, pp. 847-852.

Vafaie H & De Jong K (1993) Robust Feature Selection Algorithms. In: Proceedings of the 5th IEEE International Conference on Tools for Artificial Intelligence. Boston, USA, pp. 356-363.

Vesanto J & Alhoniemi E (2000) Clustering of the Self-Organizing Map. IEEE Transactions on Neural Networks 11(3): 586-600.

Wang G & Liao TW (2002) Automatic Identification of Different Types of Welding Defects in Radiographic Images. NDT&E International 35(8): 519-528.

Weber CA & Desai A (1996) Determination of Paths to Vendor Market Efficiency Using Parallel Coordinates Representation: A Negotiation Tool for Buyers. European Journal of Operational Research 90(1): 142-155.

Wiesinger H, Holleis G, Schwaha K & Hirschmanner F (1985) Design of CC Machines for Hot Charging and Direct Rolling. In: McMaster Symposium on Iron and Steelmaking. Department of Metallurgy and Materials Science, McMaster University, pp. 42-70.

Wilson DR & Martinez TR (1997) Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research 6(1):1-34.

Yang J & Honavar V (1998) Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems 13(2): 44-49.

Yang ZR & Chou K- (2003) Mining Biological Data Using Self-Organizing Map. Journal of Chemical Information and Computer Sciences 43(6): 1748-1753.

Appendix 1 The original features of Ruukki data

The purpose of this appendix is to list the 238 original features that composed the first dataset of Ruukki. The dataset is divided into six parts where five parts correspond to different tables in Ruukki database (see Table 25 – Table 35) and one part lists the two self-made features that were used to help the analysis (see Table 36). In Table 25 – Table 35 the last column indicates whether the feature in question was statistically analyzed at the beginning of the study and here "C" means that feature in question was used for classifying the strips into successful or retained (7 instances), "Y" means that feature was statistically analyzed (69 instances), and "N" means that feature was not statistically analyzed for various reasons.

Table 25. Features from table Primary_data_2 (85). (cont.)

Feature	Description	Analyzed
Coil_id	Identification code for a strip.	N
Timestamp	Slab discharge date and time.	N
Shift_id_char	Identification character of morning, afternoon, or night shift.	Y
Shift_id_nro	Identification number for different shifts.	Y
Analysis_nro	Number that is based on quality of slab.	Y
Analysis_version_nro	Sealing method for pass schedule calculation.	N
Material_group	Material group for pass schedule calculation.	Y
Material_class	Deformation resistance class for pass schedule calculation: based on N_strength, sealing method and material_group.	Y
N_strength	Calculated tensile strength of the strip (based on analysis results).	Y
Alloy_compensation	Alloy compensation coefficient in radiation penetration for thickness measurement.	Y
C_content	Analyzed carbon content.	Y
Si_content	Alloy silicon content.	Y
Mn_content	Analyzed manganese content	Y
P_content	Analyzed phosphorous content.	Y
S_content	Analyzed sulphur content.	Y
Al_content	Analyzed aluminum content.	Y
Nb_content	Analyzed niobium content.	Y
V_content	Analyzed vanadium content.	Y
Cu_content	Analyzed copper content.	Y
Cr_content	Analyzed chromium content.	Y

Table 26. (cont.) Features from table Primary_data_2 (85).

Feature	Description	Analyzed
Ni_content	Analyzed nickel content.	Y
Ce_content	Analyzed cerium content.	Y
Mo_content	Analyzed molybdenum content.	Y
Ti_content	Analyzed titanium content.	Y
N_content	Analyzed nitrogen content.	Y
Sn_content	Analyzed tin content.	Y
B_content	Analyzed boron content.	Y
Ca_content	Analyzed calcium content.	Y
C_eq	Carbon equivalent 3.	Y
Cm	Calculated content of carbon and manganese.	Y
Ps	Calculated content of phosphorous and sulphur.	Y
Nt	Calculated content of niobium, vanadium, and titanium.	Y
Cn	Calculated content of chromium, copper, nickel, and molybdenum.	Y
Mc	Calculated content of copper, chromium, and molybdenum.	Y
Av	Calculated content of aluminum, niobium, and vanadium.	Y
Furnace_id	Furnace identification number.	Y
Slab_storage_hours	Difference of times of slab charging and casting.	N
Slab_heating_hours	Duration of slab heating.	Y
Slab_h	Thickness of slab.	Y
Slab_l	Length of slab.	Y
Slab_w_begin	Measured width of head of slab.	Y
Slab_w_end	Measured width of tail of slab.	Y
Slab_change_begin	Starting point of change in width of slab.	N
Slab_change_place	Starting point of change in width of slab.	N
Slab_wedge_length	Length of wedge in slab.	Y
Slab_weight	Weight of slab.	Y
Slab_discharge_t_act	Actual temperature of discharged slab.	N
Rm_code	Roughing instructions.	Y
Tbar_h_tgt	Target thickness for transfer bar.	N
Tbar_w_tgt	Target width for transfer bar.	N
Tbar_l_tgt	Target length for transfer bar.	N
Tbar_lastp_t_tgt	Target temperature for last pass of transfer bar.	N
Product_code_1	Product code.	Y
Manufacturing_code	Code for manufacturing.	Y
Rolling_type	Production code.	Y
Hstr_h_tgt	Target thickness of strip.	N
Hstr_h_u_reject	Upper limit for rejecting thickness of strip.	N
Hstr_h_l_reject	Lower limit for rejecting thickness of strip.	N

Table 27. (cont.) Features from table Primary_data_2 (85).

Feature	Description	Analyzed
Hstr_flat_tgt	Target flatness of strip.	N
Hstr_flat_ul	Upper limit for flatness of strip.	N
Hstr_flat_ll	Lower limit for flatness of strip.	N
Hstr_w_tgt	Target width of strip.	N
F6_t_tgt	Target finishing temperature.	N
F6_t_min	Minimum allowed finishing temperature.	N
F6_t_max	Maximum allowed finishing temperature.	N
Surf_quality	Surface quality of strip.	Y
Hstr_uncooled_head_l	Uncooled strip length at head end.	N
Hstr_uncooled_tail_l	Uncooled strip length at tail end.	N
Cooling_strategy	Cooling strategy.	Y
Cooling_rate_tgt	Target cooling rate.	N
Cooling_1st_step_t_tgt	Temperature target for the exit of first cooling step in two step cooling.	N
Cooling_time_intermediate	Target of the intermediate time between the cooling steps.	N
Coil_t_tgt	Target temperature for coiling.	N
Coil_t_min	Minimum allowed coiling temperature.	N
Coil_t_max	Maximum allowed coiling temperature.	N
Coil_outer_diameter	Calculated outer diameter of coil.	N
Coil_bands_n	Number of bands in coil banding.	N
Coiling_strength	Tensile strength at coiling thickness and temperature.	Y
Coil_head_t_tgt	Differing coiling temperature target at head end.	N
Coil_head_l_tgt	Length of head end with differing coiling temperature target.	N
Coil_tail_t_tgt	Differing coiling temperature target at tail end.	N
Coil_tail_l_tgt	Length of tail end with differing coiling temperature target.	N
Intermediate_cooling	Intermediate cooling.	N
Customer_quality_id	Quality identification number according to specification by customer.	Y
Order_number	Number of order.	N

Table 28. Features from table Fm_prod_data_buf (100). (cont.)

Feature	Description	Analyzed
Timestamp	Time from the telegram's header.	N
Shear_time	Time of strip head reaching the front of the shear.	N
Comment_code_1	Comment code 1.	C
Comment_code_2	Comment code 2.	N
Comment_code_3	Comment code 3.	N

Table 29. (cont.) Features from table Fm_prod_data_buf (100).

Feature	Description	Analyzed
Comment_code_4	Comment code 4.	C
Reject_code_1	Automatic rejection code 1.	N
Reject_code_2	Automatic rejection code 2.	N
Reject_code_3	Automatic rejection code 3.	C
Reject_code_4	Automatic rejection code 4.	C
Reject_code_5	Automatic rejection code 5.	C
Hstr_h_dev_min_body	Minimum deviation of measured thickness at body of strip.	N
Hstr_h_dev_max_body	Maximum deviation of measured thickness at body of strip.	N
Hstr_h_dev_avg_body	Average deviation of measured thickness at body of strip.	Y
Hstr_h_dev_avg_head	Average deviation of measured thickness at head end.	N
Hstr_h_dev_avg_tail	Average deviation of measured thickness at tail end.	N
Hstr_h_min_value	Absolute minimum value of measured thickness.	N
Hstr_w_dev_min_body	Minimum deviation of measured width at body of strip.	N
Hstr_w_dev_max_body	Maximum deviation of measured width at body of strip.	N
Hstr_w_dev_avg_body	Average deviation of measured width at body of strip.	Y
Hstr_w_dev_avg_head	Average deviation of measured width at head end.	N
Hstr_w_dev_avg_tail	Average deviation of measured width at tail end.	N
F6_t_min	Measured minimum temperature after finishing mill.	N
F6_t_max	Measured maximum temperature after finishing mill.	N
F6_t_avg	Measured average temperature after finishing mill.	Y
F6_t_avg_head	Measured average temperature after finishing mill at head end.	N
F6_t_avg_tail	Measured average temperature after finishing mill at tail end.	N
Coil_t_min	Measured minimum temperature at coiler.	N
Coil_t_max	Measured maximum temperature at coiler.	N
Coil_t_avg	Measured average temperature at coiler.	Y
Coil_t_avg_head	Measured average temperature at coiler at head end.	N
Coil_t_avg_tail	Measured average temperature at coiler at tail end.	N
Tbar_t	Measured temperature at roughing mill.	Y
Tbar_t_avg_up_head	Measured average temperature before finishing at upside of head end.	N
Tbar_t_avg_up_body	Measured average temperature before finishing at upside of body of strip.	N
Tbar_t_avg_up_tail	Measured average temperature before finishing at upside of tail end.	N
Tbar_t_avg_down_head	Measured average temperature before finishing at downside of head end.	N
Tbar_t_avg_down_body	Measured average temperature before finishing at downside of body of strip.	N

Table 30. (cont.) Features from table Fm_prod_data_buf (100).

Feature	Description	Analyzed
Tbar_t_avg_down_tail	Measured average temperature before finishing at downside of tail end.	N
Tbar_t_min	Measured minimum temperature before finishing.	N
Tbar_t_max	Measured maximum temperature before finishing.	N
Active_stand_num	Number of active stands.	Y
Reject_h_l	Length of rejected thickness.	Y
Reject_profile_l	Length of rejected profile.	Y
Reject_w_l	Length of rejected width.	Y
Reject_f6_t_l	Length of rejected temperature after finishing mill.	Y
Reject_coil_t_l	Length of rejected temperature at coiler.	Y
H_class1	Thickness class 1.	N
H_class2	Thickness class 2.	N
H_class3	Thickness class 3.	N
H_class4	Thickness class 4.	N
H_class5	Thickness class 5.	N
Profile_class1	Profile class 1.	N
Profile_class2	Profile class 2.	N
Profile_class3	Profile class 3.	N
Profile_class4	Profile class 4.	N
Profile_class5	Profile class 5.	N
W_class1	Width class 1.	N
W_class2	Width class 2.	N
W_class3	Width class 3.	N
W_class4	Width class 4.	N
W_class5	Width class 5.	N
F6_t_class1	Temperature after finishing mill class 1.	N
F6_t_class2	Temperature after finishing mill class 2.	N
F6_t_class3	Temperature after finishing mill class 3.	N
F6_t_class4	Temperature after finishing mill class 4.	N
F6_t_class5	Temperature after finishing mill class 5.	N
Coil_t_class1	Coiling temperature class 1.	N
Coil_t_class2	Coiling temperature class 2.	N
Coil_t_class3	Coiling temperature class 3.	N
Coil_t_class4	Coiling temperature class 4.	N
Coil_t_class5	Coiling temperature class 5.	N
Hstr_h_max_dev_cl	Maximum thickness deviation of strip at centre line.	N
Hstr_h_avg_dev_cl	Average thickness deviation of strip at centre line.	Y
Hstr_h_min_dev_cl	Minimum thickness deviation of strip at centre line.	N

Table 31. (cont.) Features from table Fm_prod_data_buf (100).

Feature	Description	Analyzed
Hstr_location_h_max_cl	Location of the maximum thickness of strip at centre line.	N
Hstr_location_h_min_cl	Location of the minimum thickness of strip at centre line.	N
Hstr_h_avg_dev_cl_head	Average thickness deviation of strip at centre line at head end.	N
Hstr_h_avg_dev_cl_tail	Average thickness deviation of strip at centre line at tail end.	N
Hstr_coil_crown_max	Maximum crown of strip at coiler.	N
Hstr_coil_crown_avg	Average crown of strip at coiler.	Y
Hstr_coil_crown_min	Minimum crown of strip at coiler.	N
Hstr_coil_crown_avg_head	Average crown of strip at coiler at head end.	N
Hstr_coil_crown_avg_tail	Average crown of strip at coiler at tail end.	N
Hstr_h_max_rs	Maximum thickness of strip on roll change side.	N
Hstr_h_avg_rs	Average thickness of strip on roll change side.	Y
Hstr_h_min_rs	Minimum thickness of strip on roll change side.	N
Hstr_h_location_max_rs	Location of maximum thickness of strip on roll change side.	N
Hstr_h_location_min_rs	Location of minimum thickness of strip on roll change side.	N
Hstr_h_avg_rs_head	Average thickness of strip on roll change side at head end.	N
Hstr_h_avg_rs_tail	Average thickness of strip on roll change side at tail end.	N
Hstr_h_max_ds	Maximum thickness of strip on drive side.	N
Hstr_h_avg_ds	Average thickness of strip on drive side.	Y
Hstr_h_min_ds	Minimum thickness of strip on drive side.	N
Hstr_h_location_max_ds	Location of maximum thickness of strip on drive side.	N
Hstr_h_location_min_ds	Location of minimum thickness of strip on drive side.	N
Hstr_h_avg_ds_head	Average thickness of strip on drive side at head end.	N
Hstr_h_avg_ds_tail	Average thickness of strip on drive side at tail end.	N
Hstr_l_profile_g	Length of strip from profile gauge.	Y
Hstr_wedge_avg	Average of wedge of strip.	Y

Table 32. Features from table Stripheader_vw (46). (cont.)

Feature	Description	Analyzed
Coil_time	Time of strip coiling.	N
Hstr_strip_l	Length of strip.	Y
Hstr_profile_tgt	Target value for profile.	N
Hstr_profile_l_limit	Lower tolerance of strip profile.	N
Hstr_profile_u_limit	Upper tolerance of strip profile.	N
Hstr_w_l_limit	Lower tolerance of width of strip.	N
Hstr_w_u_limit	Upper tolerance of width of strip.	N
F6_roll_force	First value for rolling force of sixth stand.	Y
Op_mode_word_1	16-bit code for operation mode, word 1.	N

Table 33. (cont.) Features from table Stripheader_vw (46).

Feature	Description	Analyzed
Op_mode_word_2	16-bit code for operation mode, word 2.	N
Op_mode_word_3	16-bit code for operation mode, word 3.	N
Op_mode_word_4	16-bit code for operation mode, word 4.	N
Op_mode_word_5	16-bit code for operation mode, word 5.	N
F1_upper_grind	Grinding type of upper work roll of first stand.	N
F2_upper_grind	Grinding type of upper work roll of second stand.	N
F3_upper_grind	Grinding type of upper work roll of third stand.	N
F4_upper_grind	Grinding type of upper work roll of fourth stand.	N
F5_upper_grind	Grinding type of upper work roll of fifth stand.	Y
F6_upper_grind	Grinding type of upper work roll of sixth stand.	Y
F1_bending_force	Used rolling bending force at first stand.	N
F2_bending_force	Used rolling bending force at second stand.	N
F3_bending_force	Used rolling bending force at third stand.	N
F4_bending_force	Used rolling bending force at fourth stand.	N
F5_bending_force	Used rolling bending force at fifth stand.	N
F6_bending_force	Used rolling bending force at sixth stand.	N
F1_tension	Used tension at first stand when strip head is at sixth stand.	N
F2_tension	Used tension at second stand when strip head is at sixth stand.	N
F3_tension	Used tension at third stand when strip head is at sixth stand.	N
F4_tension	Used tension at fourth stand when strip head is at sixth stand.	N
F5_tension	Used tension at fifth stand when strip head is at sixth stand.	N
F6_tension	Used tension at sixth stand when strip head is at sixth stand.	N
F1_shift	Used axial shifting at first stand when strip head is at sixth stand.	N
F2_shift	Used axial shifting at second stand when strip head is at sixth stand.	N
F3_shift	Used axial shifting at third stand when strip head is at sixth stand.	N
F4_shift	Used axial shifting at fourth stand when strip head is at sixth stand.	N
F5_shift	Used axial shifting at fifth stand when strip head is at sixth stand.	N
F6_shift	Used axial shifting at sixth stand when strip head is at sixth stand.	N
Reject_code_word	16-bit code for rejection status.	N
Reject_a1_h	Automatic rejection code for thickness of strip.	N
Reject_a2_profile	Automatic rejection code for profile of strip.	N
Reject_a3_flatness	Automatic rejection code for flatness of strip.	C
Reject_a4_w	Automatic rejection code for width of strip.	N
Reject_a5_f6_t	Automatic rejection code for temperature of strip after sixth stand.	N
Reject_a6_coil_t	Automatic rejection code for temperature of strip at coiler.	N
Reject_scaleb_rm	Rejection code for scale breaker at roughing mill.	N
Reject_scaleb_fm	Rejection code for scale breaker at finishing mill.	C

Table 34. Features from table Rm_fault_data (1).

Feature	Description	Analyzed
Manual_comment_code_1	Possible manual comment code for a strip.	N

Table 35. Features from table Fm_reject_data_buf (4).

Feature	Description	Analyzed
Reject_code_1	Possible rejection code for a strip.	N
Reject_code_2	Possible rejection code for a strip.	N
Reject_code_3	Possible rejection code for a strip.	N
Reject_code_4	Possible rejection code for a strip.	N

Table 36. Self-made features that were used to help the analysis (2).

Feature	Description
Shift_n	Shift_id_char from table Primary_data_2 converted into number format.
Comm_rej	The amount of how many of the following features had a value other than zero: Fm_prod_data_buf: Comment_code_1, Comment_code_2, Reject_code_3, Reject_code_4, Reject_code_5 Stripheader_vw: Reject_A3_flatness, Reject_scaleb_fm

271. Wang, Lingyun (2007) The key activities of partnership development in China—a study of Sino-Finnish partnerships

272. Aikio, Janne P. (2007) Frequency domain model fitting and Volterra analysis implemented on top of harmonic balance simulation

273. Oiva, Annukka (2007) Strategiakeskeinen kyvykkyyden johtaminen ja organisaation strateginen valmius. Kahden johtamismallin testaus

274. Jokinen, Hanna (2007) Screening and cleaning of pulp—a study to the parameters affecting separation

275. Sarja, Tiina (2007) Measurement, nature and removal of stickies in deinked pulp

276. Tóth, Géza (2007) Computer modeling supported fabrication processes for electronics applications

277. Näsi, Jari (2007) Intensified use of process measurements in hydrometallurgical zinc production processes

278. Turtinen, Markus (2007) Learning and recognizing texture characteristics using local binary patterns

279. Sarpola, Arja (2007) The hydrolysis of aluminium, a mass spectrometric study

280. Keski-Säntti, Jarmo (2007) Neural networks in the production optimization of a kraft pulp bleach plant

281. Hamada, Atef Saad (2007) Manufacturing, mechanical properties and corrosion behaviour of high-Mn TWIP steels

282. Rahtu, Esa (2007) A multiscale framework for affine invariant pattern recognition and registration

283. Kröger, Virpi (2007) Poisoning of automotive exhaust gas catalyst components. The role of phosphorus in the poisoning phenomena

284. Codreanu, Marian (2007) Multidimensional adaptive radio links for broadband communications

285. Tiikkaja, Esa (2007) Konenäköä soveltavan kuituanalysaattorin ja virtauskenttäfraktionaattorin mittausten yhteydet kuumahierteen paperiteknisiin ominaisuuksiin. Kokeellinen tutkimus

286. Taparugssanagorn, Attaphongse (2007) Evaluation of MIMO radio channel characteristics from TDM-switched MIMO channel sounding

Book orders:
OULU UNIVERSITY PRESS
P.O. Box 8200, FI-90014
University of Oulu, Finland

Distributed by
OULU UNIVERSITY LIBRARY
P.O. Box 7500, FI-90014
University of Oulu, Finland

UNIVERSITY OF OULU P.O. Box 7500 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

ISBN 978-951-42-8668-1 (Paperback)
ISBN 978-951-42-8669-8 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)