Heli Koskimäki

UTILIZING SIMILARITY INFORMATION IN INDUSTRIAL APPLICATIONS
HELKOSKIMÅKI

UTILIZING SIMILARITY INFORMATION IN INDUSTRIAL APPLICATIONS

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 March 13th, 2009, at 12 noon

OUŁUN YLIOPISTO, OULU 2009
Koskimäki, Heli, Utilizing similarity information in industrial applications
Faculty of Technology, Department of Electrical and Information Engineering, University of Oulu, P.O.Box 4500, FI-90014 University of Oulu, Finland
Oulu, Finland

Abstract
The amount of digital data surrounding us has exploded within the past years. In industry, data are gathered from different production phases with the intent to use the data to improve the overall manufacturing process. However, management and utilization of these huge data sets is not straightforward. Thus, a computer-driven approach called data mining has become an attractive research area. Using data mining methods, new and useful information can be extracted from enormous data sets.

In this thesis, diverse industrial problems are approached using data mining methods based on similarity. Similarity information is shown to give an additional advantage in different phases of manufacturing. Similarity information is utilized with smaller-scale problems, but also in a broader perspective when aiming to improve the whole manufacturing process. Different ways of utilizing similarity are also introduced. Methods are chosen to emphasize the similarity aspect; some of the methods rely entirely on similarity information, while other methods just preserve similarity information as a result.

The actual problems covered in this thesis are from quality control, process monitoring, improvement of manufacturing efficiency and model maintenance. They are real-world problems from two different application areas: spot welding and steel manufacturing. Thus, this thesis clearly shows how the industry can benefit from the presented data mining methods.

Keywords: data mining, manufacturing, process data, real-world application, similarity
Acknowledgements

This study was carried out in the Intelligent Systems Group (ISG) at the Department of Electrical and Information Engineering of the University of Oulu, Finland, between the years 2004 and 2009.

I am very grateful to Professor Juha Röning and Doctor Perttu Laurinen, who supported and encouraged me throughout this work. I also thank Professor Heikki Mannila from Helsinki University of Technology and Professor Xiaohui Liu from Brunel University for reviewing this manuscript.

My colleagues from the Intelligent Systems Group also deserve my profound thanks. Special thanks go to Eija Haapalainen, who was one of my best friends from the day we began to work in the same room and helped keep me sane during these years. Also Pekka Siirtola, Henna Mörsäri and many other colleagues have been friends and not just people sharing the same workplace. I could not have dreamed of a better place to work.

This work was financially supported by Infotech Graduate School, The National Technology Agency of Finland, the European Union, the Kaupallisten ja tekniillisten tieteiden tukisäätiö and the Tauno Tönning Foundation. I gratefully acknowledge their support. In addition, I would like to thank all the partners I have cooperated with during these years.

I could not have been able to finish this thesis without my good friends. Although they have not contributed to the research part of the thesis, they have always been there for me. Especially I would like to thank Tuire, who helped me clear my head many times. My mother and father have always let me search for my own path and have encouraged me when necessary. Finally, I would like to thank my husband, Timo, who has made my life complete.

Oulu, March 2009
Heli Koskimäki
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APEN</td>
<td>Average prediction error of the neighborhood</td>
</tr>
<tr>
<td>ASSRN</td>
<td>Average of the squared standardized residuals in the neighborhood</td>
</tr>
<tr>
<td>C</td>
<td>Current</td>
</tr>
<tr>
<td>CBR</td>
<td>Case-based reasoning</td>
</tr>
<tr>
<td>F</td>
<td>Force</td>
</tr>
<tr>
<td>HWH</td>
<td>Harms + Wende GmbH &amp; Co.KG</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronic Engineers</td>
</tr>
<tr>
<td>ISG</td>
<td>Intelligent Systems Group</td>
</tr>
<tr>
<td>kA</td>
<td>Kiloamper</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge discovery in databases</td>
</tr>
<tr>
<td>knn</td>
<td>$k$ nearest neighbor</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean absolute prediction error</td>
</tr>
<tr>
<td>mm</td>
<td>Millimeter</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>robL</td>
<td>robustified Negative log-likelihood</td>
</tr>
<tr>
<td>PCB</td>
<td>Printed circuit board</td>
</tr>
<tr>
<td>SBT</td>
<td>Stanzbiegetechnik</td>
</tr>
<tr>
<td>SOM</td>
<td>Self-organizing map</td>
</tr>
<tr>
<td>V</td>
<td>Voltage</td>
</tr>
<tr>
<td>wknn</td>
<td>Weighted $k$ nearest neighbor</td>
</tr>
</tbody>
</table>
List of original publications


*Koskimäki née Junno
## Contents

Abstract
Acknowledgements 5
Abbreviations 7
List of original publications 9
Contents 11

1 Introduction 13
  1.1 Similarity utilization within this work .................................... 14
  1.2 Contributions of the thesis ............................................... 15

2 Background 19
  2.1 Data mining in industrial applications ................................. 19
    2.1.1 Data set selection .................................................. 22
    2.1.2 Case-based reasoning .............................................. 24
    2.1.3 Similarity ........................................................... 26
    2.1.4 Model adaption ..................................................... 27
  2.2 Application areas of this thesis ........................................ 27
    2.2.1 Resistance spot welding .......................................... 28
    2.2.2 Steel manufacturing .............................................. 30

3 Methods 33
  3.1 \( k \) nearest neighbors classifier .................................. 34
  3.2 \( k \)-means clustering .................................................. 34
  3.3 Self-organizing maps ................................................... 35
  3.4 Bayesian belief networks .............................................. 36
  3.5 Similarity aspect of the methods ...................................... 38

4 Utilizing similarity 39
  4.1 Quality control .......................................................... 39
    4.1.1 Data set ........................................................... 40
    4.1.2 Quality control using Bayesian networks ...................... 40
    4.1.3 Quality control using Self-organizing maps .................. 42
  4.2 Process monitoring ...................................................... 45
  4.3 Improvement of manufacturing efficiency ............................ 47
    4.3.1 Data set and features ............................................. 48
4.3.2 Process identification using self-organizing maps .......... 49
4.3.3 Extension for process identification ......................... 50
4.3.4 Ways to improve efficiency ................................ 54
4.4 Model maintenance ............................................. 56
  4.4.1 Data set and regression models ......................... 57
  4.4.2 Finding a correct moment for model update ............ 59
  4.4.3 Selecting a suitable subset for training data .......... 62

5 Summary and conclusions 67
References 68
Original publications 75
1 Introduction

In recent years the amount of digital information around us has exploded. In everyday life digital pictures, documents and music files are stored in computers. On the other hand, in industry, huge databases have been developed to store information gathered from manufacturing processes. The intent is to use this data, for example, to improve the production process or to assess product quality. However, the main question is how to handle all this data and find desired correlations from the data; for example, the human eye cannot find correlations between multidimensional data and measured quality. Thus, a computer-driven approach called data mining (also called knowledge discovery) has become an attractive research area. The term can be defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Frawley et al. 1992). Data mining research also addresses management of huge amounts of data.

In industry, different kinds of information need to be mined from gathered data. For example, the intent can be to determine the quality of separate tasks, to monitor processes or to use the information to improve the whole production process. In practice, this means research in the area is concentrated on finding solutions to specific problems. Several articles have been published concerning solutions to industrial problems using computer learning methods. They have been used in manufacturing fault detection, quality control, decision support and many other problems (Harding et al. 2006). The approaches are scattered, although due to the diversity of these tasks, a universal solution is not even possible. However, some generic principles can be adapted to several different research problems. In this thesis, an idea based on similarity of data points is utilized in diverse industrial problems. The problems not only come from two different application areas, they also vary inside the application areas. Similarity is utilized in smaller-scale problems like quality control of individual spot welding joints, but also in a broader perspective when aiming to improve the whole manufacturing process. On the other hand, the way similarity information is utilized and corresponding methods are used varies between these tasks, but the benefits of the similarity approach are clearly shown in every case.
1.1 Similarity utilization within this work

This thesis introduces how similarity information can be utilized intelligently in industrial applications. Similarity between data points and sets is shown to provide valuable information in different kinds of data mining problems. The applications introduced here are distinct, but together they clearly show the benefits of this approach. The problems covered are from quality control, process monitoring, improvement of manufacturing efficiency and model maintenance.

In many practical applications, the quality of the final product is of greatest interest (Harding et al. 2006). In addition, in most cases, actual quality cannot be measured directly, but has to be assessed using process-related data. This is an area for data mining solutions. The approach introduced in this thesis constitutes finding similar cases, studying their quality, and using this information to assess the quality of unknown data. In addition, using our approach, parameter settings that constitute low-quality products can be seen directly.

In process monitoring, similarity is utilized when developing models for controlling ongoing processes. In the model training phase, data from different process states are divided into areas comprised of similar observations whose quality is known. An optimal area is then identified, and in the control phase, the similarity of every new observation is compared with this optimal area. If the new observation is not classified as similar, a drift is noticed in the process. The similarity information is also used to enable continuation of high-quality production. In practice, this means adjustment of machining parameters based on this similarity information.

Improvement of manufacturing efficiency is also approached by utilizing similarity. In this study, data sets obtained from different processes are compared. New data are compared with data already collected from different processes, the most similar process among them is identified (process identification) and detailed information on the most similar process is retrieved from an archive. The two main benefits achieved from knowing the most similar process are: the quality of new products can be predicted and improved using the stored information of a similar process and the time needed to set up a new process can be substantially reduced by restoring process parameters that lead to high-quality results. These benefits make the overall manufacturing process more efficient, with time, money, and effort being saved from the beginning of production.

In model maintenance, information within similar data of new observations is utilized to decide if the models in use should be retrained. This is because collection of
new kinds of data and running time decrease the reliability of the models, which can in turn cause economic losses to manufacturers. Thus, having a method for deciding when to update the models becomes an important step in improving modeling in the long run. The optimal moment is sought based on estimation accuracies of the models of the most similar data points and the actual number of similar data points. If there are plenty of similar data points the allowed prediction errors are narrowed down, while if there are only a few similar data points the prediction error could be higher without affecting the need for model retraining. On the other hand, similarity information is also utilized to reduce the amount of data needed to train the models without losing interesting pieces of information. This is an important part of model maintenance, since if the amount of gathered data increases it can become infeasible to use all the data, for example, because of the memory and calculation capacities of the data mining methods. Training data selection is based on a cluster structure where similar data points constitute their own clusters. Using the cluster structure, the number of similar data points can be reduced while rarer observations are all selected into the training data.

1.2 Contributions of the thesis

This section introduces the contributions of this thesis. The previous section described how the similarity approach can be utilized in four different applications: quality control, process monitoring, improvement of manufacturing efficiency and model maintenance. The main contribution of this thesis is to show that similarity information can be utilized in diverse phases of manufacturing processes. The phases considered are from steel manufacturing to joining different steels together, and from single quality control tasks to improvement of whole manufacturing processes.

The other contributions are in novel research ideas and in the development of solutions to be applied to real manufacturing problems. These contributions are introduced next, one publication at a time. The contributions of the author of this thesis are discussed in the last paragraph.

This thesis consists of seven publications.

Publications I and II describe quality assessment of a spot welding process. In them Bayesian networks and self-organizing maps are applied for the first time to spot welding quality control. The methods were chosen because they do not make strong assumptions about the distribution of the data and thus adjust easily to real data (the conditional independence assumption often made with Bayesian networks is not con-
sidered to be of high importance from the application point of view). In addition, the results of the methods are simple to understand even without any profound knowledge of the methods themselves, which is desirable when aiming to develop systems for use in industrial production lines.

Publication III introduces a novel idea of using self-organizing maps to discover process drift in spot welding. When a drift in a manufacturing process is noticed online, certain counteractions can be performed immediately, enabling better quality during the whole process. In addition, the publication also introduces a way to perform the counteractions.

Publications IV and V deal with process identification. In process identification the idea is to gather data from different manufacturing processes, and when collecting new data, to retrieve a similar data set from a database and utilize the information of the retrieved process. After a similar process is identified, two different steps can be taken: (1) restore the process parameters leading to high-quality products and substantially reduce the time needed to set up a new process or (2) seek the quality prediction and improvement methods already developed for the similar process. In addition, a specific similarity measure was developed to decide if the retrieved process is actually similar enough that the information can be used. In the spot welding industry, no comparable approach aiming to improve the whole manufacturing process could be found. Thus, the process identification task introduced in this thesis is considered an important novel step for the spot welding industry.

Publications VI and VII concentrate on maintenance activities for models. For example, collection of new data and the passing of time decrease the reliability of models. This can be solved by retraining models at optimally selected intervals. Publication VI introduces a novel approach to finding the optimal time for model updating. Despite an extensive search by the author, no previous research reported in the area of model updating in a real application, where updating was developed without a periodic approach, was found. Thus, the developed approach can be assumed to be a significant contribution to the area of model efficiency continuum. Publication VII describes a new procedure for selecting a suitable subset of instances for model training so that the reduction does not negatively affect the accuracy of the model. In addition, a novel clustering method called coarse clustering is introduced in publication VII. Both publications demonstrate the effectiveness of the proposed method by using them with predictive regression models developed to optimize steel manufacturing process settings beforehand.
In publications where the author of this thesis is presented as the first author, the study and discussion are mainly direct contributions of the author, while the studies were planned and in the writing phase the contents of the publications were revised by the other authors. In two of the publications the author of this thesis was not the first author. However, in publication I the study was done by the author, although the results were mainly reported by P. Laurinen. In publication III the basic idea was presented by the author, although the study and discussion were done by the other authors.
2 Background

This chapter introduces the background for this thesis. It contains a short overview of the main activity in the area of data mining and of applications in the manufacturing industry in the 21th century; areas related the most to this thesis. In addition, the sub-tasks of data mining related to the thesis are discussed. Data set selection, which is very important in every practical application, is covered in subsection 2.1.1. Research done by utilizing previous cases (case-based reasoning) and the background of similarity are introduced in subsections 2.1.2 and 2.1.3, respectively. In addition, subsection 2.1.4 includes a short discussion about the adaptation capability of models. Section 2.2 introduces work done in two application areas: spot welding and steel manufacturing.

2.1 Data mining in industrial applications

As computers became more common and the amount of data gathered increased, an idea for finding suitable patterns from data using a computer-driven approach was introduced. In business, data mining was thought to give a competitive advantage (Frawley et al. 1992). In a workshop of the U.S. National Science Foundation held in 1990, data mining was positioned as one of the most promising research areas of the 1990s (Silberschatz et al. 1990). It has not failed these expectations; data mining conferences are held worldwide, hundreds of research papers are published in many journals and books have been written, not only on actual data mining methods, but also on the background support of these methods.

The term data mining was at some time thought of as a pattern recognition step of the KDD process (knowledge discovery in databases) (Fayyad et al. 1996), which is depicted in Figure 1. However, nowadays the term ‘data mining’ is understood to cover the whole process and can be applied to any data source, not just databases. The synonym ‘knowledge discovery’ is also used to describe the whole process (Kurgan & Musilek 2006). The methods used in the data mining process consist of summarization, classification, dependency modeling, clustering, link analysis and sequence analysis (Wang 1999).
In manufacturing, data mining solutions are used to model, classify and make predictions in numerous applications (Harding et al. 2006). The application areas covered with data mining approach are (Wang 2007):

1. optimization of manufacturing yield,
2. quality control,
3. monitoring and diagnosis,
4. preventive machine maintenance,
5. manufacturing process control,
6. process planning and scheduling,
7. material requirement planning,
8. manufacturing system modeling,
9. safety evaluation,
10. assembly selection,
11. learning robotics.

The first six application areas are most related to this thesis. The main aim in every manufacturing application is to make a profit, meaning that optimization of manufacturing yield is the driving force behind every application. Of the individual application areas, quality control is approached in publications I and II, and monitoring and diagnosis as well as preventive machine maintenance and manufacturing process control are covered in publication III. Item 6 is related to publications IV and V. Publications VI and VII
are not directly related to any of the applications, although in them also, optimization of manufacturing yield is the final aim. Thus, the study carried out in publications VI and VII can be considered separated from the application outline.

Computer-aided quality control in industrial applications aims to check the quality of products in cases where quality measurement is expensive, unreasonable for humans or even impossible without breaking the final product. The machine learning approaches most commonly used consist of neural networks, fuzzy logic, evolutionary algorithms and different classification methods. In addition, the application areas tackled are spread out to widely cover different areas of manufacturing. Studies have been conducted to estimate the quality of complex electronic products, for example, sound speakers (Melin & Castillo 2002) and PCB boards (Yang & Tsai 2002). Also quality in different areas of manufacture in basic industry are covered; many scientific publications discuss the areas of pulp and paper (Achiche et al. 2006), steel (Wiltshi et al. 2000) and wood (Silven et al. 2003) manufacturing. On the other hand, quality assessment methods have been developed to control certain stages of production. For example, the quality of individual spot welding joints (Xinmin et al. 2007) and the quality of the painting process (Filev 2002) are evaluated individually in car assembly. In addition, many visual inspection methods have been developed to monitor the quality of products on an assembly line, where human concentration can be distracted from the task from time to time (Piuri et al. 2005).

In monitoring and diagnosis the main aim is to keep the manufacturing process efficient in the long run by diagnosing system failures and avoiding process disorders. It is also related to preventive machine maintenance and process control. The solution approaches are divided into three different categories: rule-based, case-based and model-based diagnostic systems (Hur et al. 2006). In the rule-based solution faults are diagnosed using certain rules that tell the correlation between monitored equipment condition and each possible fault. In case-based systems historical maintenance information is utilized. In model-based systems different mathematical, statistical and machine learning methods are used to diagnose faults based on the structure and properties of the systems used. In data mining the solution is usually a combination of available information and machine learning methods. For example, in the semiconductor industry faults can be detected using the $k$ nearest neighbors method. The training data consist of normal situations, and faulty situations are detected if the distance from a new case to the normal cases is too big (He & Jin 2007). There are also studies where the system is based on sensors and programmable logic controllers, and the data set for fault
identification is gathered from them (Hou et al. 2003). In addition, plenty of effort has been used for very specific aspects of process monitoring, for example, Rehorn et al. (2005) review more than 100 scientific articles on solely tool condition monitoring systems in conventional cutting operations, including drilling, turning, end milling and face milling.

Manufacturing process planning and scheduling are widely approached using case-based reasoning methods. There are studies where artificial intelligence is utilized to plan and schedule whole assembly lines, for example, in car manufacturing (Gusikhin et al. 2007). However, this review emphasizes how the processes are initialized to produce high-quality products from the beginning. Thus, the research done to find optimal setup parameters for production machinery is introduced in more detail. Sadoyan et al. (2006) sought suitable parameters for the spray tooling process in tool and die making using if/then rules. Tsao (2008) studied the correct setup value for the thrust force of a step drill based on the step angle, the stage ratio, the feed rate and the spindle speed. In addition, genetic algorithms have been used to select the optimal parameters of a mold oscillation system in the continuous casting process of steel manufacturing (Bhattacharya et al. 2007).

One feature is common to all the introduced manufacturing applications. The data mining solutions are developed for specific problems and it is unlikely that the solutions could be adapted easily to different applications. Thus, from the industrial point of view, there is still plenty of research to be done. In addition, there are new application areas where data mining is not yet utilized. Even in the traditional application areas there is still a need for deeper technical solutions. The model’s maintenance activities and manufacturing efficiency improvement introduced in this thesis are examples of such deeper solutions.

2.1.1 Data set selection

As data mining methods are becoming more common, more interest is focused on the background of these methods. Data preparation is considered to be essential for successful mining applications (Pyle 1999). For example, the learning capability of machine learning models depends heavily on the features given to the models.

The term ’feature identification’ is used to describe the search for suitable features. It can be divided into two sometimes overlapping categories: feature extraction and feature selection. In feature extraction the features affecting the problem are searched
Features can be, for example, certain individual measurement values and/or, as in many cases where the data consist of measurement signals, certain statistical or geometrical features that represent the signals (Zhang et al. 2006). These features can be reworked, for example, using PCA to uncorrelate them or using feature selection methods to choose a subset of features that are the most suitable for the problem. For example, Haapalainen et al. (2006) introduced different methods for selecting the most suitable feature set. The aim can be to minimize the size of the feature set (makes the algorithms used work faster and/or the memory requirements smaller) or improving prediction accuracy. In many cases the smallest possible feature set that does not significantly decrease prediction accuracy is chosen (Dash & Liu 1997).

When the available data set is considered as a matrix where the features are the columns of the matrix and the observations constitute the rows, a distinct approach comparable to feature selection is easy to rationalize. In feature selection the problem of an overwhelming amount of data is tackled by removing meaningless columns from the data set. In some cases the data set reduction has to be performed by deleting insignificant rows from the data set (Liu & Motoda 2002). When the suitable features are found but the number of available data points increases, it is difficult to use all the data in data mining applications. The procedure in which the number of data points is reduced is called instance selection, though instance selection is a less-treated research problem than feature identification.

Although the question of instance selection is tackled in several articles from the classification point of view (Brighton & Mellish 1999, 2002, Jankowski & Grochowski 2004) very few articles consider instance selection with other kinds of problems. In addition, most of the methods introduced in classification-oriented articles are not applicable when class labels are not available. The few instance selection studies found that are adaptable to a case where class labels are unavailable have solved the problem using genetic algorithms (Kim 2006) or with an approach based on tuning random sampling. For example, a data set is selected using random sampling and tuned by replacing random observations of the selected data with observations left out if the replacement improves the model’s accuracy (Wilson & Martinez 2000). However, in most cases this approach would be impractical and highly time-consuming. Nevertheless, approaches based on k-means clustering have given promising results in the classification area (Rokach et al. 2003, Altınçay & Ergün 2004) and they can also be adapted to a case where class labels are not available.
2.1.2 Case-based reasoning

Utilization of similar data sets is recognized as an important aspect in many different research areas. In many of these studies the problem is referred to as case-based reasoning. Case-based reasoning (CBR) is an approach where new problems are solved using prior experience (Lopez de Mantaras et al. 2006). It can also be considered a method of data mining.

The first steps in case-based reasoning were taken by Roger Schank in 1982 and Janet Kolodner, a researcher in Schank’s group. Kolodner developed the first actual case-based reasoning system, CYRUS, in 1983. Elsewhere, Bruce Porter brought different aspects to the development of case-based reasoning in 1986 and the system developed using his research was called PROTOS (Aamodt & Plaza 1994).

Aamodt (Aamodt & Plaza 1994) has described CBR as a cyclical process of four Rs:

1. RETRIEVE the most similar case or cases
2. REUSE the information and knowledge in that case to solve the problem
3. REVISE the proposed solution
4. RETAIN the parts of this experience likely to be useful for future problem solving.

Figure 2 shows the task-method decomposition of case-based reasoning. Although in the decomposition the tasks are considered subtasks of case-based reasoning, many of these tasks themselves are actually independent research areas, and their introduction would not be sensible from the point of view of this thesis. Thus, the discussion concentrates on actual use cases in industrial applications.
The case-based reasoning approach has been used in many different industrial applications. Case-based reasoning is recognized to be a functional aid in many industrial design tasks. For example, to reduce design costs, Boyle et al. (2004) used CBR for fixture design. Woon et al. (2005) used CBR in the design of conveyors, for example, to decide on suitable input parameters. Case-based reasoning systems have also been found suitable in oil and gas well design (Kravis & Irgang 2005). In addition, the case-based reasoning approach is used for process planning of automotive panels (Chen et al. 2005).

A second area in industry where case-based reasoning methods are found effective is quality improvement. In the metal industry case-based reasoning is used in quality designing, where the most similar case is retrieved using the $k$ nearest neighbors method combined with fuzzy logic (Kim et al. 2004), and in defining a suitable temperature value for blowing control using only the $k$ nearest neighbors method (Kim et al. 2005). The $k$ nearest neighbors method is also used as a retrieval method for fault diagnosis of industrial robots (Olsson et al. 2004) and to control water quality and billing in the water supply industry (Saward 1999). In addition, a case-based reasoning system is used to control an electric furnace for reduction of copper from slag (Moczulski & Szulim 2004).
2.1.3 Similarity

In case-based reasoning the solutions are based on similar previous events. Similarity information can offer huge benefits in other data mining methods, also. To specify similarity information, the points to be concentrated on are the features used to define similarity, the similarity measures and the domain knowledge available. It should be noted that their selection depends on the method used and the form of data set available.

When considering similarity from the point of view of suitable features, surface similarity is used in many cases. In surface similarity the features are presented as vectors or attribute-value pairs. Other commonly used methods are $k$ nearest neighbors (the number of cases to be retrieved is predefined, but the actual similarity of the cases to the target problem varies) or different window methods (the number of cases to be retrieved varies, but the similarity to the target problem is inside a predefined threshold) (Lopez de Mantaras et al. 2006).

When using the $k$ nearest neighbors method for retrieval of similar cases, the features are assigned different weights. It has been shown that classification accuracy can be improved by limiting the number of prototypes and by weighting features (Skalak 1994). Different methods have been developed to automatically decide the proper weights for features, for example by Cardie (1993), Moore & Lee (1994), Skalak (1994) and Aha & Bankert (1994). Many of the automated weight assignment methods properly give low weights to irrelevant features. However, by using performance feedback the results improve compared with methods that do not use feedback (Wettschereck & Aha 1995).

Feature weighting is also with other methods. Múnoz-Avila & Hüllen (1996) introduced a method where two cases are considered similar if a certain predefined number of features match regardless of what the features are. However, because some of these features are more important than others, they have to be weighted differently.

In addition, the similarity measures that are used should be considered. In the $k$ nearest neighbors method, for example, an Euclidean or Mahalanobis distance measure is used to define the nearest neighbors (Mitchell 1997). These measures give a numeric value to similarity. Also ordinal and Boolean values for similarity can be used (Osborne & Bridge 1996). The similarity measure is normally chosen according to the problem (Zhang et al. 2006).

Another very important aspect in similarity is deciding when the most similar is not similar enough. In other words, when the new data point or data set is so different
compared with the data already gathered that it could be categorized as a novel event. This task is usually known as novelty detection.

Selection of a method for novelty detection depends on the statistical properties of the data set. Parametric methods assume that the data distribution is Gaussian and the problem can be modeled statistically based on data means and covariance (Markou & Singh 2003). However, their effectiveness depends more on the fact that the data can only support simple decision boundaries such as linear or quadratic (Hastie et al. 2001). Also non-parametric approaches such as nearest neighbor, Parzen density and string matching methods can be used for novelty detection (Markou & Singh 2003).

### 2.1.4 Model adaption

In some studies, for example (Gabrys et al. 2005), model adaptation has been considered as the model’s ability to learn behavior in areas from which information has not been acquired. However, in this thesis, adaptation of the model is considered to be the ability to react to time-dependent changes in the modeled causality. In practice, model adaptation means retraining the model at optimally selected intervals. However, because the system has to adapt quickly to a new situation in order to avoid losses to the industry, periodic retraining, used in many methods ((Haykin 1999), (Yang et al. 2004)), is not considered to be the best approach. Moreover, there are also disadvantages if retraining is done unnecessarily. For example, extra work is needed to take a new model into use in the actual application environment. In the worst case, this can result in coding errors that affect the actual accuracy of the model. In spite of extensive literature searches, studies that would be comparable with the approach were not found. Thus, it can be assumed that the approach is new, at least in an actual industrial application.

### 2.2 Application areas of this thesis

This section introduces the background of the application areas where the approach of this thesis is utilized. The next subsections explain what has been done in the areas using soft computing and why data mining is so important. Although the introduced studies from the application areas have not separately tackled the question of similarity, it generally forms the basis of every modeling task. In model training it is assumed that the available data represent the whole data space and thus the capability of the models to estimate new data points is related to the existence of similar data points in
the training data.

2.2.1 **Resistance spot welding**

Resistance welding is the most common way to join metal objects together. The most common type of resistance welding, developed by Professor Elihu Thomson, is called resistance spot welding. It is widely used, for example, in electrical and automotive industries, because it is effective and economical. In addition, the spot welding process is easy to automate (Rautaruukki 2002).

Figure 3 shows an example of spot welding. Two metal sheets are welded together. Electrode force is applied to hold the sheets tightly together, and electrical current flows through the electrodes and the material. The resistance of the material being welded is much higher than the resistance of the electrodes. Thus, enough heat is generated to melt the metal. The pressure on the electrodes forces the molten spots in the two pieces of metal to unite, forming the final spot (nugget).

![Fig 3. Principle of spot welding (Publication V).](image)

The most important thing in this application area is to assure that the spots are produced with good quality. Because the actual quality can only be measured using destructive testing, meaning breaking the actual product, it can be done only randomly, and there is a need for data mining solutions to ensure quality. Also important to manufacturers is how to produce good-quality spots in cases where the product is made for the first time or there is no clear information on how to produce it.
Research on computational quality assessment techniques in the field has concentrated on estimating the quality of welding by using neural networks, regression analysis and mathematical methods. The studies have utilized different features extracted from data. The variation of resistance over time (dynamic resistance pattern) has been an important explanatory variable in many of the studies. Artificial neural network and regression models have been generated based on the dynamic resistance pattern by, for example, Aravinthan et al. (2001) and Cho & Rhee (2002, 2004). Cho & Rhee (2002) compared regression analysis with neural networks, which demonstrated the superior accuracy of the neural network estimator. Unfortunately, the sample consisted of only 60 measurements, and the leave-one-out method was used to measure the estimator’s performance, which limits the significance of the conclusions. Also the new approach of Cho & Rhee (2004), where a Hopfield network was used to classify a new experiment into five different classes (two of them consisted of unsuccessful welds), lacks confidence, because the test set was comprised of only ten experiments. In addition, dynamic resistance has been used to monitor welding processes online (Garza & Das 2001). However, their data set was narrow, consisting of only 40 experiments. Studies using other input variables include approaches involving neural networks with tip force, the number of weld cycles, the weld current and the upslope current (Ivezic et al. 1999). Also pre-weld information such as angular misalignment and fit-up faults of the electrodes are used (Brown & Schwaber 2000). Common to all these studies is that the quality control methods are developed merely to estimate quality, while using similarity information also the parameter values affecting quality could be seen.

Although in many studies the quality control of spot welding is approached using data mining solutions, no reported approaches aiming to improve the whole manufacturing process could be found. Thus, the process identification task, introduced in this thesis, where the characteristics of a sample measured from a new process are compared with information gathered from previous processes to find a similar process, is considered an important step for the spot welding industry. This is because the retrieved information can be used to return the process parameters leading to high-quality joints (the time needed to set up a new process can be substantially reduced) or to predict the quality of new welding spots using the stored information of a similar process.
2.2.2 Steel manufacturing

Steel is the most important structural material used in industry; there is plenty of iron (Fe) in the ground, steel is easy to manufacture, cheap, easily moldable, has high strength and it is possibility to affect the mechanical properties of products by changing material compositions, process variables and mechanisms (Nevalainen 1970). Steel manufacturing is a multi-step process where liquid steel is cast into steel slabs that are then rolled into steel. Figure 4 shows an example of how steel slabs are processed after coming out of the furnace.

Fig 4. Steel manufacturing (hot strip mill of Ruukki Steel) (Elsilä & Röning 2002).

Data mining inspired widespread research in the area of steel manufacturing. Different properties affecting steel quality are studied separately. For example, tensile strength and yield strength are modeled by joining information on the predicted mean and dispersion, aiming to solve beforehand the rejection probabilities in qualification tests (Juutilainen & Röning 2004) or the actual limits inside which the strength value will be (Juutilainen & Röning 2006). Research by Tamminen et al. (2008) concentrated on impact toughness estimation. The surface defects of steel have also been considered. For example, Haapamäki et al. (2005) studied the scale defects of steel and Jia et al. (2004) examined seam defects.

On the other hand, models have been developed to control the intermediate phases of manufacturing, for example, to predict the post roughing mill temperature of steel slabs (Laurinen et al. 2001) and to monitor scheduled working times in relation to observed times using multiagent technologies, making it possible to notice faulty situations rapidly (Fischer et al. 2004). In addition, optimization of coolant bow settings in the spray cooling zone has been studied for possible casting process improvements (Filipic & Robic 2004).
As it was shown, data mining has been utilized in many different sectors of steel manufacturing. However, the studies have concentrated on basic modeling of certain tasks, while in this thesis the emphasis is on model maintenance. In the basic modeling task the similarity aspect is not a necessary approach, but in model maintenance similarity information proved its usability. While the model maintenance task introduced in this thesis is a novel approach, it is not surprising that similarity information has not been used this way previously. No comparable studies in which the ultimate goal is to maintain steel quality prediction models automatically or even to improve model accuracy in the long run were found. Thus, also from the steel manufacturing point of view, the steps taken in this thesis are valuable and welcome.
3 Methods

When considering data mining research, different machine learning methods are widely used. Some of the methods are so-called black box methods, meaning that after the training the method itself does not carry information on which are the most important features, how the different features correlate with each other or whether there are similar observations among the data. The last of these items is the most interesting point within the scope of this thesis. Although every model can be thought to benefit from similarity, for example, a model’s prediction capability can be considered to correlate with the existence of similar data points in the training data, there are a smaller number of models that do not lose similarity information. Thus, in this thesis, the methods used were chosen not only based on their capability to solve the problem in question, but also to introduce how similarity information can be brought forward.

From the basic applicability point of view, the methods used in this thesis, Bayesian belief networks, a self-organizing map, a \( k \) nearest neighbor classifier and \( k \)-means clustering, were chosen because they do not make strong assumptions about data distribution and therefore adjust to real-world problems where data distributions are not known. An exception to this is the conditional independence assumption of Bayesian networks, but the assumption is not considered to have a great effect on the adjustability of the method with real-world data. Another advantage of the methods is that the results they give are easy to interpret without a deep understanding of the underlying principles of the methods. Thus, it is no coincidence that \( k \)-means clustering, the \( k \) nearest neighbor classifier and a simplified version of the Bayesian networks (naive Bayes) were ranked by the IEEE International Conference on Data Mining in 2006 among the top 10 most influential data mining algorithms in the research community (Wu et al. 2008).

As the discussion on the similarity point of view of the methods assumes an understanding of the basic principles of the methods, the methods are introduced in the next subsections and the similarity discussion is postponed until the last subsection. The basic discussion about the methods is mainly focused on the principles of the methods and the applications the methods are used for.
3.1 \( k \) nearest neighbors classifier

The \( k \) nearest neighbors classifier (knn) is a method for which it has been proven that, when \( k=1 \), the classification error is maximally twice the optimal Bayesian probability of error (Cover & Hart 1967). However, the idea of \( k \) nearest neighbors classification is quite simple: a data point is classified into the class where most of its \( k \) nearest neighbors belong (Fix & Hodges Jr. 1951). The nearest neighbors are defined using, for example, the Euclidean distance measure (Mitchell 1997).

Different modifications have been made to the basic algorithm. In the weighted \( k \) nearest neighbors method (wknn) the class is defined using weighted sum voting, meaning that the class is decided using not only the number of neighbors from certain classes, but also their distance to the data point to be classified. The closer the neighbors are, the more weight is given to the class they represent (Hechenbichler & Schliep 2004). Also, scaling the features differently is a widely used method for increasing the weight of the most significant features in the classification.

The \( k \) nearest neighbors classifier and its variations are widely used in real-world applications. They have been used, for example, to categorizing texts (Wang & Wang 2007), to recognize human-performed daily activities based on triaxial accelerometer and heart rate data (Pirttikangas et al. 2006) and to reduce the number of false alarms in intrusion detection systems (Law & Kwok 2005). In industry, the \( k \) nearest neighbors method has been applied to detect broken bars in asynchronous motors (Ondel et al. 2006) and to detect surface defects on production wafers (Hunt et al. 2000). The effectiveness of the \( k \) nearest neighbor method has also been noted when monitoring semiconductor process tools, especially plasma etching (Chamness 2006).

3.2 \( k \)-means clustering

The \( k \)-means clustering method was developed to divide a given data set into a certain prefixed number of clusters, \( k \). It aims to minimize the distance inside the clusters and maximize the distance between clusters by minimizing the squared error function

\[
V = 
\sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2,
\]

(1)

where \( S_i, i = 1, \ldots, k \) are the \( k \) clusters and \( \mu_i \) are the cluster centroids and also the means of points \( x_j \in S_i \). This is done using an iterative training sequence. The \( k \) cluster
centroids are chosen randomly, after which every data point is attached to a cluster with the closest centroid point. The new centers of the clusters are then calculated and the centroid value is updated. The process is carried on iteratively until a point is reached where no changes to cluster centroids happen or a certain number of iterations have been done (Bishop 2006).

In the introduction of this chapter it was stated that the k-means method does not make any assumptions about the distributions of the data even though the basic version assumes that the different clusters can be separated by treating them as normally distributed. However, the statement was based on the fact that by using different weighting schemes and distance measures, the k-means method can be made to forget these assumptions (Wu et al. 2008). For example, by using categorical weighting for some variables or the mahalanobis distance as a distance metric, the form of the clusters can be changed.

The actual application areas of the k-means clustering algorithm are also widespread. It has been applied to face recognition (Hadid & Pietikäinen 2004) and to recommend new music tunes to people based on their previous music downloading (Kim et al. 2007). In addition, because k-means clustering is a data partition method, the solutions are often combined with other data mining methods. For example, it has been applied with neural networks to discover faults in satellite attitude determination systems (Cai et al. 2007) and with genetic algorithms in anomaly intrusion detection (Marimuthu & Shanmugam 2008). In industry, k-means clustering has been applied, for example, to reduce state space complexity in supply networks of automotive production systems, after which control rules can be learned more straightforwardly (Doring et al. 2007). In semiconductor wafer fabrication process control, k-means have been used to preprocess the data before applying genetic algorithms and support vector machines (Huang & Chen 2006).

### 3.3 Self-organizing maps

A self-organizing map (SOM) is an unsupervised neural network method that visualizes high-dimensional data in a low-dimensional (typically two-dimensional space). It was developed by Professor Teuvo Kohonen in 1981. The SOM presents the statistical dependencies of high-dimensional data in the form of geometric figures. This is done by keeping the topologic and metric relations of the two-dimensional space as close as possible to the relations of the initial high-dimensional space.
Self-organizing maps are widely used and more than 5000 scientific articles have been based on them already before 2001 (Kaski et al. 1998, Oja et al. 2003). They have been used, for example, in health monitoring (Tamminen et al. 2000), in financial analysis (Kiviluoto & Bergius 1998) and in speech recognition (Kohonen et al. 1996). Self-organizing maps have been combined with machine vision systems, for example, to develop wood surface inspection methods (Niskanen et al. 2001) and to detect human faces from pictures (Ikeda et al. 2002). Several areas of industry have noticed the effectiveness of the method. For example, self-organizing maps are used to model a steel production process and to study a continuous pulp digester (Alhoniemi et al. 1999). Many process monitoring approaches also rely on self-organizing maps (Kohonen et al. 1996).

A SOM is usually formed of neurons on a regular low-dimensional grid with a hexagonal or rectangular lattice. The neurons are model vectors \( m_i = [m_{i1}, m_{i2}, \ldots, m_{in}] \), where \( n \) is the dimension of the input space. Training is done by choosing a data sample \( x \) and finding the closest model vector \( m_c \) (the best-matching unit). When the best-matching unit is found, it and its topologically closest neighbors are updated with the equation

\[
m_i(t+1) = m_i(t) + \alpha(t)h_c(t)[x(t) - m_i(t)],
\]

where \( \alpha(t) \) is the learning rate factor (a decreasing function of time) and \( h_c(t) \) is the neighborhood kernel centered on the winner unit \( c \). Training continues by choosing a new data sample and iterating the updating equation (Kohonen 2001).

### 3.4 Bayesian belief networks

Bayesian belief networks were developed to model dependencies between variables. They have been applied to diverse problems, for example, failure analysis of inject printers (Lee 2001) and sensor planning for mobile robot localization (Zhou & Sakane 2002). The quality of C++ programs (Masoud et al. 2004) and text-to-speech systems (Goubanova 2001) are also studied using Bayesian belief networks.

A Bayesian belief network is a directed acyclic graph that consists of nodes and links between them. The nodes represent variables and the links describe conditional probabilities between them. It is applied mostly to discrete variables. Figure 5 shows an example of a Bayesian network. Node \( A \) represent a discrete variable, which can have states \( a_1, \ldots, a_n \). Every state has a probability \( P(a_i) \) and the probabilities of the all
states of variable $A$ add up to one.

**Fig 5. Bayesian belief networks.**

In addition, nodes $A$ and $B$ in Figure 5 are considered parents of $X$ and nodes $C$ and $D$ represent children of $X$. The probability distribution of $X$ is conditional on that of its parents $A$ and $B$ and its children $C$ and $D$. In other words, the probability of occurrence of a certain state of $X$ is dependent on the states of its parents and children (assuming there is information available about their states). However, the states of the children and parents affect a different amount. Using Bayesian networks, the biggest posterior probabilities can be calculated for every variable (Duda et al. 2001).

A naive Bayes classifier is a simplified method of Bayes’ theorem based on the conditional independence assumption that all attributes are independent, given the value of the class variable. Although the assumptions simplify the problem, the results of the method are in many cases even surprisingly good (Hastie et al. 2001, Hand & Yu 2001). It has actually been shown that in some cases the naive Bayes classifier outperforms more sophisticated methods, even when there are dependencies between variables (Domingos & Pazzani 1997). In addition, it is very easy to construct. It does not acquire any iterative parameter estimation and thus can be easily applied to huge data sets (Wu et al. 2008). In practical applications the naive Bayes approach is used, for example, in automatic video surveillance systems (Hongeng et al. 2000), to filter spam messages (Androutsopoulos et al. 2000) and for text classification (Eyheramendy et al. 2003).
3.5 Similarity aspect of the methods

From the similarity point of view, the methods were chosen because they do not lose similarity information, but instead similarity is one of the driving forces of the methods. The methods were introduced in order of their simplicity compared with the similarity aspect, while the most straightforward method was introduced first.

The $k$ nearest neighbor method was already mentioned in section 2.1.3 to be one of the most common methods used in similarity research. The similarity aspect of the method is quite obvious. Basic similarity is usually measured as distance between data points, and what the $k$ nearest neighbor classifier decides is based on the closest observations, which are naturally the most similar ones. Similarity and k-means clustering are also highly related in that the observations are clustered based on their distances to each other, indicating that the most similar observations are in the same or neighboring clusters.

Self-organizing maps are trained to preserve topologic and metric relations as well as possible, meaning that even in the two-dimensional plane the most similar map cells are close to each other. The SOM was built as a visualization tool and is a more sophisticated method compared to knn and k-means, but it also relies on distance information in the training phase. In addition, when using the trained map to classify new observations, the best matching unit (the most similar map cell) is searched for and the decision is made based on this unit.

Utilization of similarity is not as straightforward in Bayesian belief networks. Distance information is not used in the training phase, but instead the similarity aspect is based on the structure of the method. Dependencies are estimated for discrete variables, while the variables constitute different classes. These classes contain observations that are similar to each other in relation to the variable. For every variable there are distinct classes, but in the modeling phase the dependencies are estimated separately for every class combination. Therefore, these class combinations contain only observations that are similar to each other, and similarity information is preserved also with Bayesian networks.
4 Utilizing similarity

This chapter introduces the main contributions of this thesis. The advantages of utilizing similarity information are presented clearly in four different applications in the manufacturing industry. Section 4.1 introduces the work done for quality control of welding spots (publications I and II). Section 4.2 considers process monitoring (publication III), while section 4.3 discusses how to improve manufacturing efficiency from the point of view of spot welding (publications IV-V). The last section, 4.4, describes methods developed for model maintenance; a case study was done for the steel manufacturing industry (publications VI and VII).

4.1 Quality control

The main task here was to investigate how similarity information can be used to predict the quality of welding spots in a welding process. Quality control was approached with two different well-known methods: Bayesian networks and self-organizing maps. They were chosen because they adapt easily to real industrial data: self-organizing maps do not make any assumptions about the distribution of the data and although Bayesian networks often make assumptions to simplify data, this does not significantly affect the adjustability of the method. In addition, the results given by the methods are easy to interpret, which was considered an important feature when aiming to implement methods for actual welding production lines. Another very important aspect of the methods was that, beside their given estimate, they preserve similarity information, in which case feature sets constituting unsuccessful events can be discovered and avoided in future production.

In both methods observations are grouped with similar observations, in Bayesian networks using classes and in self-organizing maps by arranging them into the same or close map elements. When the quality of a new observation is predicted, the most similar observations are searched (the corresponding classes or map elements). The quality of the most similar cases is studied and this information is used to estimate the quality of unknown data points. The similarity information can also be used when searching for the feature sets of unsuccessful events.
4.1.1 Data set

The data used in the study were comprised of measurements from welding tests done at Voest-Alpine Transport Montage Systeme GmbH, Austria. The data set contained observations from 192 experiments in which two metal sheets were welded together with a resistance spot welding machine. Each of the observation sets contained measurements of current, compression force and voltage signals recorded during the welding and the diameter of the weld measured after welding (the diameter must exceed 4 mm to represent a good joint). The signals were measured at intervals of 0.04 milliseconds and each of them was composed of about 7000 values.

Naturally, it was not reasonable to use all these data points to study the methods, thus suitable features were extracted from the signal curves. The extracted features are introduced in detail in the next subsections.

4.1.2 Quality control using Bayesian networks

The first approach was to define the effects of different signal curves on the quality of welding joints by using Bayesian networks. The study was done together with P. Laurinen and also reported as an example in his thesis on a top-down approach to creating and developing data mining solutions (Laurinen 2006). The idea of Bayesian networks is to model conditional probabilities between variables and represent them as easily understandable rules.

In the study, the features extracted were the quartile and median values of the signals. Thus, every measured signal could be presented as three values constituting altogether nine features when all three signals were treated. These continuous values were further processed by dividing them into classes. The distributions of the continuous values were used when creating the class structure. An example of the approach can be seen in Figure 6 a), which shows a plot of the lower quartiles of the voltage for all the observations. In addition, the diameter values were also divided into classes, where class 1 contained all the unsuccessful welds (Figure 6 b)). This new data set was then used to create Bayesian networks.
Different correlations between the diameters and some of the extracted features could be clearly seen from the created Bayesian networks. For example, using only three determining features, the median of the voltage (V_med), the upper quartile of the current (C_uq) and the upper quartile of force (F_uq), the results showed which class divisions lead to unsuccessful welds. Some of the most interesting class probabilities of the welding spot diameter are shown in Table 1. The first columns in the table show the classes of the features, the probabilities of the welding spot belonging to a certain class are given in the columns marked with D and the last column shows how many of the observations belonged in the configurations.

Table 1. Probabilities from the Bayesian network. V_med = voltage median, C_uq = current upper quartile, F_uq = force upper quartile, D = welding spot diameter, #obs = number of observations in the configuration.

<table>
<thead>
<tr>
<th>Variables and their classes</th>
<th>Welding spot diameter</th>
<th>#obs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V_med</td>
<td>C_uq</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Fig 6. Division into classes for a) voltage and b) welding spot diameter (Publication I).
It can be seen from the first rows, (1, 1, 3) and (1, 2, 3), of Table 1 that with the values situated in these classes, it is highly probable that the diameter of the welding spot will be unsuccessful. Although in these cases only single observations were available, it is highly probable that in future cases the same assumptions will hold. In rows (2, 1, 2), (2, 1, 3), (2, 2, 2) and (2, 3, 2) the probabilities are more equally distributed amongst all the classes, but more information can be obtained by using more determining features. The rest of the configurations listed in the table show high probabilities of the welding spots belonging to good-quality classes, which is also very important information.

Although the results in the study are not adjusted exactly to give the best results, for example, features extracted in further studies could give more straightforward results, the basic idea and its usability in this kind of problem is obvious. Clear rules on when unsuccessful welding spots occur or may occur make it possible for manufacturers to avoid settings constituting them and in this way improve the overall quality of the welding process.

### 4.1.3 Quality control using Self-organizing maps

The second approach to quality control was based on self-organizing maps. SOMs perform clustering of feature patterns, where similar patterns form a cluster characterized by a representative pattern. The clusters are presented as a two-dimensional map (U-matrix), with the map location distances reflecting the similarity of the clusters. In addition, representative patterns can be drawn for every feature used to train the map individually so that the nodes at same places correspond with each other. The map arrangement allows direct interpretation and understanding of the interrelations between the features and their relationship to the quality.

The study was mainly based on the means of ten equally long parts of the continuously measured signals that were chosen as features. The legends used in the study are: \(c = \) current, \(f = \) force and \(v = \) voltage, and the numbers 1, \ldots, 10 mark the means of the respective tenths of the signal divided into ten parts of equal length. The study was divided into three steps:

- Reducing the feature set (SOM trained with all the features and the quality measure),
- Studying the correlation between quality and the signal features (SOM trained with a reduced feature set and the quality measure),
- Using the map to predict unsuccessful welds (SOM trained without the quality measure, using only 80 percent of the data).
In the first step the correlations between the features were studied. If the map formation was similar for two or more features, only one of them was left to represent the effect of the features in the second step. In practice, this meant a reduction from 31 features to 10 features.

The map in Figure 7 shows the effects of this reduced feature set on the diameter of welding spots. A comparison of the last feature map of diameters with the other maps shows, for example, that small compression force value has a positive effect on diameter. This can be seen through the large values of diameter and small values of force occurring in the bottom right corner. Also the largest values of voltage \( v_3 \) occur in that corner. When comparing these with the relations found using Bayesian networks in subsection 4.1.2, it can be seen that the correlations hold. For example, in Table 1 the smallest values of force correlated positively with the size of the welding spot.

On the other hand, the small values of diameter appear at the top of the feature map. The features that affect the diameters at the top are the smallest values of voltage, \( v_4 \) and \( v_5 \). The smallest values of current in the upper right corner and the largest values of force in the upper left corner also contribute to the small values of diameter. Furthermore, these results are consistent with the results in the previous subsection.
To estimate the quality prediction capability of the SOMs, the data were divided into
training data (80 percent of the data) and testing data (20 percent). The map was trained
with the training data; diameter information was not used as a training parameter. How-
ever, the diameter labels were assigned afterwards to the map elements representative
of the curves that belong to the corresponding cluster (Figure 8 a)). An area that con-
tains all the unsuccessful welds of the training data can be found in the figure (marked
with grey). When looking for the best-matching unit on the map for testing data points
and adding the diameter labels, it can be seen (Figure 8 b)) that all the unsuccessful
welds of the testing data are placed in the grey area derived from Figure 8 a). Only
three successful welding spots were misinterpreted, but they lie on the same nodes as
the successful welds in Figure 8 a). Depending on the area of application, the division
into successful and unsuccessful welds can be changed. If it is important that all the
unsuccessful welding spots are found, the division shown in the figure can be used, but
if some of the unsuccessful spots can be misinterpreted, the grey area can be smaller.

Fig 8. The trained map when the uncorrelated features of signals and diameter are
used as the training parameters (Publication II).
4.2 Process monitoring

In process control the basic idea was to monitor the manufacturing processes online. A self-organizing map (SOM) was taught with observations from different process states and an area of optimal process states was discovered. When a new process was launched the new observations were mapped into the SOM. If a drift from the optimal area was identified while the process was ongoing, necessary changes were made instantly to enable continuation of high-quality production. The functionality of the system was verified with spot welding data.

In this study similarity was utilized in both phases: process monitoring and process adjustment. The self-organizing map was used to divide the training data into nodes comprised of similar observations. For each node, the quality (poor, average, good) was estimated and an optimal area was decided from the map. In the control phase the best matching node (the most similar) was searched for every new observation. If the most similar node was not from the optimal area, the process was adjusted by changing the machining parameters based on the distance and angle between these two areas.

The principle of the system in a spot welding case can be seen in Figure 9. The current and voltage signals gathered during every separate welding event were merged into a resistance signal. The resistance signal was preprocessed and suitable features were extracted from it using principal component analysis. The self-organizing map and the developed drift detector were used to find the distance and angle to the optimal area and the converter was used to adjust the current and pressure based on this information.
Figure 10 shows an example of a trained self-organizing map. The qualities assigned to each node are color-coded, meaning that red and green shades mark better quality. The optimal area found is marked as a rectangle and the center of it is marked as a white circle. When a drift is detected, like in this case when the best matching node (marked as yellow circle) is pretty far from the optimal area, the distance and the angle to the optimum center can be calculated.
Fig 10. Self-organizing map with process optimum area marked as rectangle. The white circle marks the optimum center, while the yellow circle marks the winner node for a new welding event (Publication III).

The results showed that this approach is applicable in a real-world problem. The overall production duration of high-quality welding joints was lengthened and the system was easily set to the appropriate manufacturing environment.

4.3 Improvement of manufacturing efficiency

In this task similarity information was used to identify similar welding processes. The ultimate aim of the system was to gather vast amounts of information from different manufacturers into a database from which the most similar process will be sought. After the similar process is identified, its detailed information is retrieved from the archive. The information can be retrieved for two distinct purposes: (1) to restore the process parameters leading to high-quality joints and substantially reduce the time needed to set up a new process or (2) to seek the quality prediction and improvement methods already developed for the similar process. Thus, the developed system makes the overall manufacturing process more efficient while time, money and effort are saved from the beginning of production.
The principle of this study can be seen in Figure 11. First the data from a new process are gathered and pre-processed. The new data are compared with the data already stored in the database and the most similar process is identified among them (process identification). After that the similarity measure between the new data set and the most similar process is calculated. If the processes are classified as similar based on the similarity measure, the database restores the process parameters leading to high-quality joints or the methods optimal for quality control and improvement. The study on the database build-up is reported in (Tuovinen et al. 2007).

**Fig 11. The principle of the study (Publication V).**

### 4.3.1 Data set and features

In the study, the data set was provided by two welding machine manufacturers: Harms + Wende GmbH & Co. KG (HWH) (Harms+Wende www page) and Stanzbiegetechnik (SBT) (Stanzbiegetechnik www page). The HWH tests were done on materials commonly used in the automotive industry, while SBT concentrates on welding thin materials used in electronic components.

The data set contained altogether 3879 experiments from 20 processes (11 from HWH and 9 from SBT, and they are marked as HWH1-HWH11 and SBT1-SBT9). The experiments were welded using varying welding parameters, and each experiment consisted of measurements of current and voltage signals recorded during the welding process. The processes were divided into several configurations. The term 'configu-
ration’ refers to a set of experiments (usually 15) welded with unchanged parameters. For example, the experiments measured with current values of 8.0 kA, 7.2 kA and 6.4 kA while the other parameters remained unchanged constituted three different configurations. The parameters that varied between the configurations were current, electrode force, electrode wear and fitting.

The measured signal curves contained plenty of oscillatory motion and a pre-heating section, and they were hence preprocessed before further manipulation. The pre-heating parts of the curves were cut off, leaving only the signal curves recorded during the actual welding phase. In addition, the curves were smoothened using the Reinsch algorithm (Reinsch 1967, 1971). After pre-processing, suitable features were extracted from the signal curves. As in section 4.1.3, the averages of ten parts of equal length were used as features.

### 4.3.2 Process identification using self-organizing maps

The first steps to process identification were taken using self-organizing maps. The visualization capability of the method was considered a benefit to actual application usage. In addition, SOMs were used to find the correct initialization parameters, which could be taken into use after identifying the closest process.

At this point the data set consisted of five SBT processes, and ten means were calculated for the resistances of the signal curves. The classification ability of the self-organizing maps was verified by dividing the data into training data and test data. The map was trained with 80 percent of the available data, and in Figure 12 a) the observations of five test series are labeled according to corresponding map nodes with the numbers 1-5. However, the division in the lower part of the map only points out the differences inside processes 1 and 2, which are thought to be similar processes, and it is therefore not considered an important division. In addition, the processes of an independent test set were classified correctly using the map.

The results showed that self-organizing maps can be used for process identification.
After the study, more data were archived and the accuracy of the self-organizing maps decreased. Thus, the visualization capability had to be forgotten and a more accurate method of classification had to be sought.

### 4.3.3 Extension for process identification

Haapalainen *et al.* (2005) carried out a comparative study of different parametric and non-parametric classification methods for spot welding process classification. In their study, the \( k \) nearest neighbor method was found to be the most suitable for the process identification task. Nevertheless, the approach was based on identifying the most similar process and a more accurate examination of the similar process was not executed. This means that, although a similar process is retrieved from a database using the method, how can we be sure it is similar enough to be used to give suitable initialization parameters and quality control methods. Thus, the emphasis was shifted not only to process identification but also on actual similarity between processes.
In this study the knn method was used to find the most similar process. In addition, an extension of the method was developed to find out if the most similar process could be classified as similar, because it was recognized that some of the new processes could be so dissimilar from the ones already stored in the database that leaving out the distance information would cause serious misclassifications. The approach was also reported in Junno et al. (2005).

The idea of similarity was approached by developing two boundaries, taking into consideration that different processes were divided into several configurations (a set of experiments welded with unchanged parameters). In practice, the first boundary calculated the average distance to $k$ ($k = 5$) nearest neighbors inside the configurations, while the second one expressed the average distance to $k$ nearest neighbors between two neighboring configurations. The configurations by which the actual second boundary value was calculated differed only with regard to the initialized current, because current was the only parameter that varied between the configurations in every process and can be assumed to vary in future processes. The final threshold, called the similarity measure, was a combination of these boundaries.

A simplified example of the boundaries is presented in Figure 13. Actually, the data set was 20-dimensional, but the idea is presented here using a two-dimensional figure. The white circles refer to measurements from two neighboring configurations of a process. The black circles refer to test points 1-3. The test points demonstrate the new measurements to be classified, and the grey dashed circles around them mark the boundaries.
The final similarity measure combines the information of the boundaries and provides a more straightforward result to be used in industry. The final form of the similarity measure was obtained by comparing these two boundaries with the average distance from the data points of a new test set to the $k$ nearest neighbors of the closest matching process. If the new distance is marked as $d$ and the boundaries are marked as $b_1$ and $b_2$, then the similarity measure $\text{sim}$ is

$$\text{sim} = \frac{d - b_1}{b_2 - b_1}.$$  \hspace{1cm} (3)

Now the thresholds for the similarity measure can be formulated more concretely. If the similarity value is smaller than zero, the new test set is considered to be precisely the same process as the closest matching process. If the similarity measure is between zero and one, the closest process is considered similar, and knowledge of that process can be used as tentative information of the new test set. If the similarity value is greater than one, the closest matching process is classified as dissimilar.

In this study, each configuration of different processes was successively considered as a new data set (information of the process to which the configuration belongs is not taken into account) and compared with the information stored in the database, making the setting resemble the actual operation of the system. To avoid repetition, only three of the processes are introduced in detail. Table 2 shows the results obtained from the
first five configurations of the processes. The abbreviations C1-C5 represent different configurations. In addition, the first number of cells indicates the percentage of data points of a configuration classified as each process, while the second number is a value of the similarity measure.

Table 2. Classification of the first five configurations of a) HWH2, b) SBT3 and c) SBT4.

<table>
<thead>
<tr>
<th>Config</th>
<th>a) HWH2</th>
<th>b) SBT3</th>
<th>c) SBT4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HWH3</td>
<td>HWH4</td>
<td>SBT9</td>
</tr>
<tr>
<td></td>
<td>%</td>
<td>SM*</td>
<td>%</td>
</tr>
<tr>
<td>C1</td>
<td>100</td>
<td>0.49</td>
<td>100</td>
</tr>
<tr>
<td>C2</td>
<td>93.3</td>
<td>0.37</td>
<td>6.7</td>
</tr>
<tr>
<td>C3</td>
<td>13.3</td>
<td>0.32</td>
<td>86.7</td>
</tr>
<tr>
<td>C4</td>
<td>100</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>100</td>
<td>0.21</td>
<td></td>
</tr>
</tbody>
</table>

*Sm = Similarity measure

The HWH2 case in Table 2 indicates that most of the configurations are classified with 100 % accuracy as HWH3. The configurations C2 and C3 are exceptions to this. For example, 93.3 percent of the experiments of configuration C2 are classified as process HWH3 and 6.7 percent as process HWH4. In this case, the similarity measure is also better for HWH3. However, most of the experiments in configuration C3 are classified as HWH4, while the similarity measure is significantly smaller for HWH3. Hence, the user needs to decide which should be considered more important, a small value of the similarity measure or a large classification percentage. However, the system developed in this study shows both similarity and percentage, which enables the user to make the decision by applying information on both processes and choosing the most suitable alternative.

In the SBT data set, the similarity measure showed that some of the most similar processes were farther than others. An example of two processes known to be approximately the same is shown in Table 2 b). Processes SBT3 and SBT9 are similar to each other, but the SBT9 process was welded one year later. Table 2 b) shows the classification results for SBT3. It can be seen that the similarity measures for all the
configurations are smaller than one, and the processes are therefore classified as similar. An opposite example can be seen in Table 2 c) for process SBT4, for which there are no similar processes stored in the database. The similarity measures give the same information. The configurations of process SBT4 are classified as SBT9, but the similarity measures are much larger than one. These two processes are classified as dissimilar using the similarity measure; this was also the desired result.

### 4.3.4 Ways to improve efficiency

When the most similar process is retrieved and it has been determined that it is similar enough, the benefits of the system introduced in Figure 11 can be achieved. The simplest benefit is that the correct initialization parameters can be restored from the database, which reduces the time needed for the machine setup. In addition, different data mining methods developed for quality control or process drift estimation of the most similar process can be employed as such or with small modifications, enabling certain delivery of high-quality end products to the customers. These benefits are introduced in more detail in the next subsections.

**Restore of correct initialization parameters**

When a similar process is found, the database is searched and the parameters leading to high-quality joints are restored. This approach can substantially reduce the setup time of a new process. In practice, even experienced welding machine users must find suitable parameters through trial and error. Naturally, this means spending more time and money, while the system developed here can restore the parameters faster and more automatically. The ultimate aim of the system is to gather vast amounts of information from different manufacturers into a database from which suitable initialization parameters can be searched independent of the manufacturer. Thus, when a manufacturer begins to weld new parts (expertise needed to select the suitable parameters is not found inside the factory) the initialization parameters already used in a different factory can be found from the database.
Restore of quality control and process monitoring methods

Quality control and improvement methods that already exist can be used for a new process if the process that is found is classified as similar. A simple example of quality prediction can be seen in Figure 14. The average quality of the \( k \) closest data points of a new experiment is assigned as a quality value of the experiment while the results are introduced configuration-wise (= as a mean of approximately 15 observations). In Figure 14 a) the most similar process is classified as similar, while in Figure 14 b) the most similar process is classified as dissimilar. It can be seen that the measured and predicted values in similar case correspond well to each other. Naturally, when the similarity measures are poor the results are remarkably worse. In that case, the quality information of the closest matching process cannot be used, but because the similarity measure is known, this mistake would not be made.

![Fig 14. Average measured and predicted quality values of experiments by different configurations of a) similar processes and b) dissimilar processes (Publication V).](image)

Figure 15 shows an example how detection and visualization of process drifts can be utilized for a new process. The basic idea for one process was introduced in section 4.2. In this case, the self-organizing map is trained using information from a similar process, while the map is also used to detect and visualize the process drifts of a new process. Figure 15 a) shows the U matrix of a self-organizing map of the restored process, while Figure b) shows a labeled map where the labels demonstrate successful (marked as 1) and unsuccessful (marked as 0) welding experiments. It can be seen that the unsuccessful welds are located in the upper and lower left corners. The grey area in Figure 15 b) is used to mark the optimal process state area where the highest quality welding spots are situated.
Fig 15. Self-organizing map for process drift detection and visualization (Publication V).

The labels demonstrated in Figure 15 c) are labels of the quality of the new process, which are assigned to the map elements of the restored process. The same optimal, grey area is also used for process drift detection in this case. It can be seen that the unsuccessful welding experiments of both processes are situated in approximately the same areas. This means a previously trained map of a similar process can help improve the quality of the welding spot of a new process.

4.4 Model maintenance

The application area where the functionalities of the methods introduced in the previous sections were verified was spot welding. In the model maintenance case the results are from the steel industry. Already now, development engineers control mechanical properties such as yield strength, tensile strength, and elongation of the metal plates beforehand on the basis of planned production settings. The solution is achieved using regression models introduced by Juutilainen & Röning (2006). However, acquisition of
new data and the passing of time decrease the reliability of the models, which can bring economic losses to the plant. Thus, maintenance of the models emerges as an important step in improving modeling in the long run. The next subsections describe how to find the correct moment for a model update and how to select suitable instances for model fitting when all the available data cannot be used.

In model maintenance, similarity information is used in two ways. In a model update, information from previous similar data points is used to decide the optimal time for model retraining. Information on the estimation accuracy of the models for similar data points and on the number of similar data points is combined to form a limit, called the exception limit. This limit is then used in deciding on the need for model retraining. In the training data selection, similarity information is used to reduce the amount of data needed to train the models without losing interesting pieces of information. In practice this is done in the selection by weighting fewer data points having plenty of similar observations in the data space. The similar data points are discovered using clustering, which divides the data points into clusters based on their similarity.

### 4.4.1 Data set and regression models

The data for this study were collected from the production database of Ruukki steel works between July 2001 and April 2006. The whole data set consisted of approximately 250,000 observations. Information was gathered on element concentrations in actual ladle analyses, normalization indicators, rolling variables, steel plate thicknesses, and other process-related variables (Juutilainen et al. 2003). The observations were gathered during actual product manufacturing. The volumes of the products varied, but if there were more than 500 observations from one product, the product was considered a common product. Products with fewer than 50 observations were categorized as rare products.

In the studied prediction model, the response variable used in the regression modeling was the Box-Cox-transformed yield strength of the steel plates. The Box-Cox transformation was selected to produce a Gaussian-distributed error term. The deviation in yield strength also depended strongly on input variables. Thus, the studied prediction model included separate link-linear models for both mean and variance

\[
y_i \sim N(\mu_i, \sigma_i^2) \\
\mu_i = f(x_i^T \beta)
\]
\[ \sigma_i = g(z_i^2 \tau). \quad (4) \]

The length of the parameter vector of mean model \( \beta \) was 130 and the length of the parameter vector of variance model \( \tau \) was 30. The input vectors \( x_i \) and \( z_i \) included 100 carefully chosen non-linear transformations of the 30 original input variables; for example, many of these transformations were products of two or three original inputs. The link functions \( f \) and \( g \) that were used were power transformations selected to maximize the fit with data. The results are presented in the original (nontransformed) scale of the response variable (Juutilainen & Röning 2006).

**Goodness measures for regression models**

The data set is constituted of different amounts of observations from different products; there could be thousands of observation from the most common products and only a few from the rarest products. Thus, a normal average of absolute prediction error would skew the regression models toward even small improvements in the prediction accuracy of the most common products. To prevent skewness, two different goodness measures were introduced where the prediction accuracy of the observations of the rarest products was emphasized.

The accuracy of the models was measured in the property prediction case using the weighted mean absolute prediction error

\[ \text{MAE} = \frac{1}{\sum_{i=1}^{N} w(i)} \sum_{i=1}^{N} w(i) |y_i - \hat{\mu}_i|. \quad (5) \]

In the variation model case, a robustified negative log-likelihood was employed to take into account the variance

\[ \text{robL} = \frac{1}{\sum_{i=1}^{N} w(i)} \sum_{i=1}^{N} w(i) \left( \log(\hat{\sigma}_i^2) + \rho \left( \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} \right) \right). \quad (6) \]

Here, the function \( \rho(\cdot) \) is a robust function; this study employed

\[ \rho(t) = \begin{cases} t, & \text{when } t \leq 25 \\ b^2, & \text{when } t > 25, \end{cases} \quad (7) \]

which truncates the squared standardized residuals if the standardized residual is below -5 or above +5.

Two different methods were used to define the weights, \( w(i) \). They were chosen to reflect the usability value of the models. In the first goodness criterion (Goodness 1)
the weights \( w(i) \) were defined by product, meaning the weight of the observations of a product could be at most as much as the weight of \( T \) observations. Let \( T_i \) be the number of observations that belong to the same product as the \( i \)th observation. Then the weight of observation \( i \) is

\[
  w(i) = \begin{cases} 
    1, & \text{when } T_i \leq T \\
    \frac{T}{T_i}, & \text{when } T_i > T. 
  \end{cases}
\] (8)

Here the value of \( T = 50 \), meaning if there are more than 50 observations of a product, the weight is scaled down. The second goodness criterion (Goodness 2) was formed to take only rare observations into account. Thus, only the cases for which there were only fewer than 30 previous observations within a distance 0.9 or a distance 1.8 were included, but the weight of the latter was double (equation 9)

\[
  w(i) = \begin{cases} 
    1, & \text{when } \{\#x_j | |x_i - x_j| < 0.9 \& j < i\} < 30 \\
    2, & \text{when } \{\#x_j | |x_i - x_j| < 1.8 \& j < i\} < 30 \\
    0, & \text{else.} 
  \end{cases}
\] (9)

### 4.4.2 Finding a correct moment for model update

In this study the correct moment for a model update was approached using information from previous cases. The update was based on the average prediction errors and the average of the squared standardized residuals of deviations of similar past cases. The similar past cases inside a certain distance limit constituted a neighborhood and the average values inside the neighborhood were obtained using equations 10 and 11.

Koskimäki et al. (2007) reported a preliminary study where the update was considered only from the point of view of model prediction errors.

The average prediction error of the neighborhood (= \( APEN \)) was calculated as the distance-weighted mean of the prediction errors of observations belonging to the neighborhood:

\[
  APEN = \frac{\sum_{i=1}^{n} [(1 - \frac{d_i}{\max(d)}) \cdot \hat{e}_i]}{\sum_{i=1}^{n} (1 - \frac{d_i}{\max(d)})},
\] (10)
where

\[ n = \text{number of observations in a neighborhood}; \]
\[ \varepsilon(i) = \text{the prediction error of the } i\text{th observation of the neighbourhood}; \]
\[ d_i = \text{the Euclidian distance from the new observation to the } i\text{th observation of the neighborhood}; \]
\[ \text{max}(d) = \text{the maximum allowed Euclidian distance between the new observation and the previous observations in the neighborhood (}=3.5). \]

The average of the squared standardized residuals in the neighbourhood (\(=\ ASSRN\)) was obtained using:

\[
ASSRN = \frac{\sum_{i=1}^{n} \left[ (1 - \frac{d_i}{\text{max}(d)}) \cdot \frac{\varepsilon_i^2}{\sigma^2} \right]}{\sum_{i=1}^{n} (1 - \frac{d_i}{\text{max}(d)})},
\]

(11)

where

\[ \sigma = \text{deviation achieved from the regression model}. \]

The optimal time for a model update was obtained by using two different limits: the exception limit and the update limit. The exception limit was exceeded if the value of \(APEN\) was too high in relation to the size of neighborhood or the value of \(ASSRN\) was too high or low compared with the average. The update limit was formed to merge the information of exceptions. The limit was defined as being exceeded if the sum of the exceptions within a certain time interval was too high. Whenever the update limit was exceeded the property prediction and deviation models were re-fitted.

Tables 3 and 4 show the goodness criteria when the models are updated using this approach. To compare the results, the criteria are presented in three different cases: predicted with a newly updated model, with a model updated in the previous iteration step, and with a model that was not updated at all. The results show that, although the differences between the new model and the model from the previous iteration step are not in every case very big, the update improves the prediction in each of the steps. In addition, in some steps it can be seen that even a small addition of training data improves accuracy significantly. Naturally, the benefit of the model update is obvious when the results of the updated model and the model without an update are compared.
Table 3. Means of absolute property prediction errors with the goodness criteria.

<table>
<thead>
<tr>
<th>Step size</th>
<th>With new model goodness 1</th>
<th>With previous model goodness 1</th>
<th>Without update goodness 1</th>
<th>With new model goodness 2</th>
<th>With previous model goodness 2</th>
<th>Without update goodness 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12228</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18932</td>
<td>10.48</td>
<td>11.01</td>
<td>10.70</td>
<td>13.20</td>
<td>14.15</td>
<td>14.03</td>
</tr>
<tr>
<td>5626</td>
<td>10.93</td>
<td>11.07</td>
<td>11.25</td>
<td>12.94</td>
<td>13.11</td>
<td>13.10</td>
</tr>
<tr>
<td>17636</td>
<td>12.48</td>
<td>12.57</td>
<td>12.97</td>
<td>16.48</td>
<td>16.60</td>
<td>18.23</td>
</tr>
<tr>
<td>743</td>
<td>19.94</td>
<td>20.01</td>
<td>25.72</td>
<td>57.18</td>
<td>57.48</td>
<td>71.18</td>
</tr>
<tr>
<td>3772</td>
<td>12.77</td>
<td>14.75</td>
<td>17.77</td>
<td>21.10</td>
<td>36.00</td>
<td>49.63</td>
</tr>
<tr>
<td>13533</td>
<td>12.47</td>
<td>12.62</td>
<td>13.68</td>
<td>16.34</td>
<td>17.03</td>
<td>22.37</td>
</tr>
<tr>
<td>43338</td>
<td>11.78</td>
<td>11.97</td>
<td>12.51</td>
<td>13.66</td>
<td>13.98</td>
<td>15.61</td>
</tr>
<tr>
<td>14831</td>
<td>12.78</td>
<td>13.26</td>
<td>13.77</td>
<td>21.05</td>
<td>20.98</td>
<td>25.55</td>
</tr>
<tr>
<td>21397</td>
<td>12.45</td>
<td>12.70</td>
<td>14.26</td>
<td>22.43</td>
<td>23.84</td>
<td>36.60</td>
</tr>
<tr>
<td>622</td>
<td>16.20</td>
<td>18.28</td>
<td>25.39</td>
<td>73.18</td>
<td>99.29</td>
<td>156.71</td>
</tr>
<tr>
<td>Mean</td>
<td>11.99</td>
<td>12.26</td>
<td>12.79</td>
<td>15.43</td>
<td>16.08</td>
<td>18.10</td>
</tr>
</tbody>
</table>
4.4.3 Selecting a suitable subset for training data

The previous section discussed how to recognize the optimal moment to update models when the amount of available data increases. This section introduces a procedure for selecting suitable training data instances for this update step.

Although computers are getting more efficient, there are still problems in training data mining methods with all the data gathered. Memory or calculation capacities are not sufficient enough because the sizes of data sets are also growing. To tackle this problem, a five-step instance selection algorithm was developed for selecting suitable instances from the whole data set for regression model fitting:

1. Perform coarse clustering (Algorithm 1)
2. Split coarse clusters whose size is greater than $V$ using k-means clustering
3. Select all the instances from non-partitioned coarse clusters

Table 4. Robustified negative log-likelihood values with the goodness criteria.

<table>
<thead>
<tr>
<th>Step size</th>
<th>With new model goodness 1</th>
<th>With previous model goodness 1</th>
<th>Without update goodness 1</th>
<th>With new model goodness 2</th>
<th>With previous model goodness 2</th>
<th>Without update goodness 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>12228</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36703</td>
<td>6.29</td>
<td>6.30</td>
<td>6.30</td>
<td>6.87</td>
<td>6.92</td>
<td>6.92</td>
</tr>
<tr>
<td>18932</td>
<td>6.13</td>
<td>6.27</td>
<td>6.19</td>
<td>6.57</td>
<td>6.76</td>
<td>6.72</td>
</tr>
<tr>
<td>5626</td>
<td>6.22</td>
<td>6.23</td>
<td>6.29</td>
<td>6.52</td>
<td>6.52</td>
<td>6.56</td>
</tr>
<tr>
<td>17636</td>
<td>6.38</td>
<td>6.39</td>
<td>6.46</td>
<td>7.15</td>
<td>7.13</td>
<td>7.36</td>
</tr>
<tr>
<td>6104</td>
<td>6.32</td>
<td>6.42</td>
<td>6.53</td>
<td>6.62</td>
<td>6.90</td>
<td>6.84</td>
</tr>
<tr>
<td>743</td>
<td>8.02</td>
<td>8.06</td>
<td>10.66</td>
<td>15.84</td>
<td>16.11</td>
<td>29.56</td>
</tr>
<tr>
<td>3772</td>
<td>6.36</td>
<td>6.94</td>
<td>8.58</td>
<td>7.16</td>
<td>11.58</td>
<td>23.04</td>
</tr>
<tr>
<td>13533</td>
<td>6.43</td>
<td>6.45</td>
<td>6.84</td>
<td>7.15</td>
<td>7.20</td>
<td>10.13</td>
</tr>
<tr>
<td>43338</td>
<td>6.35</td>
<td>6.38</td>
<td>6.47</td>
<td>6.67</td>
<td>6.71</td>
<td>7.33</td>
</tr>
<tr>
<td>14831</td>
<td>6.55</td>
<td>6.65</td>
<td>6.94</td>
<td>7.47</td>
<td>7.52</td>
<td>11.13</td>
</tr>
<tr>
<td>21397</td>
<td>6.40</td>
<td>6.45</td>
<td>7.13</td>
<td>7.17</td>
<td>7.54</td>
<td>15.39</td>
</tr>
<tr>
<td>622</td>
<td>6.58</td>
<td>6.76</td>
<td>11.47</td>
<td>9.65</td>
<td>11.66</td>
<td>70.85</td>
</tr>
<tr>
<td>Mean</td>
<td>6.36</td>
<td>6.41</td>
<td>6.61</td>
<td>6.91</td>
<td>7.05</td>
<td>8.28</td>
</tr>
</tbody>
</table>
4. Select instances from k-means clusters using equation 13
5. Select the rest of the instances using random sampling

In coarse clustering the data set is divided into clusters using a certain predefined distance limit so that two different clusters cannot contain any observations within this limit (Algorithm 1). After coarse clustering, k-means clustering is performed on the largest clusters.

**Algorithm 1** Coarse clustering algorithm

```
for each $x_i$ in data do
    if $\exists$ one cluster: $\text{dist}(x_j, x_i) \leq d$ & $x_j \in$ cluster then
        $\text{cluster} \leftarrow x_i$
    else
        if $\exists$ several clusters: $\text{dist}(x_j, x_i) \leq d$ & $x_j \in$ cluster then
            $\text{merge these clusters} \leftarrow x_i$
        else
            $\text{create new cluster} \leftarrow x_i$
        end if
    end if
end for
```

After clustering has been performed, the actual data selection phase follows. The instances are selected by calculating an information value for model fitting (measured on the basis of the location and spread of the cluster) for each cluster and then the number of selected observations is determined in the proportion of the cluster’s information values.

Let $C_1, C_2, \ldots, C_q$ denote the clusters resulting from the application of coarse clustering and k-means algorithms and let $c_1, c_2, \ldots, c_q$ be the corresponding cluster centroids. Let $n_i = \#\{x_j | x_j \in C_i\}$ denote the number of observations included in the $i$th cluster. Then the information contents of the $i$th k-means cluster can be measured by

$$J_i = \exp(a_1 \sigma_i + a_2 m_i)$$

(12)

where $\sigma_i = \sqrt{(1/n_i) \sum_{x_j \in C_i} ||c_i - x_j||}$ is the within-cluster deviation, $m_i = \max_j ||c_i - c_j||$ is the distance to the closest cluster centroid and $a_1, a_2$ are tuning constants. Small coarse clusters are valuable for model fitting because they are isolated from other input space regions that contain data. Thus, all observations from coarse clusters that were
not partitioned using k-means are included in the selected training data. The number of selected instances, $S_i$, of the partitioned clusters is obtained using the equation:

$$S_i = \frac{N \cdot J_i}{\sum_{j=1}^{q} J_j};$$

(13)

where $q$ is the number of k-means clusters and $N$ is the number of observations that still need to be selected after the observations of the coarse clusters have been chosen.

The goodness of the training data selection was studied by comparing the prediction accuracies of the models trained with the limited data chosen with our approach (50,000/200,000, clust), with limited data chosen randomly (50,000/200,000, rs) and with the whole data set (reference, 1-200,000, all). The prediction accuracy of the models was studied for the remaining observations (ca. 45,000). In addition, the results of our method were studied for three different cluster structures (clust 1, 2 and 3). They are introduced to show that although k-means clustering does not necessarily find the optimal solution in every run, the use of our approach to instance selection minimizes the problem.

Because the results varied during different instance selection runs using our method and using random sampling, the selection phase was repeated ten times. Using goodness criterion 1 for the property and prediction models, the results are slightly better than those obtained from random sampling, while both ways were approximately as good as the reference model. However, when considering the rarest cases (Goodness 2) and comparing the results obtained from the worst instance selection run, our approach outperformed random sampling (Table 5).

| Table 5. Best and worst goodness 2 values and standard deviation for the property and deviation models of different runs using our method and with random sampling. Smaller values mean better results and references are 22.64 and 7.49. |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | MAE             | RobL            |                 |                 |
|                 | Min  | Max  | Std  | Min  | Max  | Std  |                 |
| clust 1         | 21.48 | 22.35 | 0.28 | 7.37 | 7.55 | 0.05 |                 |
| clust 2         | 21.55 | 22.15 | 0.20 | 7.41 | 7.54 | 0.04 |                 |
| clust 3         | 21.72 | 22.59 | 0.28 | 7.38 | 7.72 | 0.10 |                 |
| rs              | 22.00 | 24.52 | 0.79 | 7.36 | 8.55 | 0.37 |                 |
For example, when considering the application-specific working limits, which are a combination of the property and deviation measures, it can be seen that the prediction accuracy achieved using random sampling can be almost nine percent smaller than in the reference case (goodness 2, MAE: 22.64 and goodness 2, RobL: 7.49). On the other hand, when comparing the prediction accuracies achieved using our approach, better results can be seen. In clust 1 and 2 the combined prediction capability of the property and deviation models is always better than in the reference case. Although in clust 3 the prediction capability is weaker than in the reference case, the difference is only 0.5 percent, which is not considered a significant drop.

From the application point of view, the instance selection procedure should be consistent in every run. In addition, it is desirable that it would not weaken the model’s ability to predict rare steel grades, meaning the method is able to generalize at the boundaries of the training data. Based on the results, it seems that these can be fulfilled using our approach. Thus, our approach is suitable for the instance selection task.
5 Summary and conclusions

This thesis addresses the issue of how similarity information can be utilized in industrial applications. Nevertheless, in general, every modeling task can be thought to utilize similarity: the models are trained with data that is assumed to represent the data space and thus the estimation capability of the models with new data points is related to the existence of similar data points in the training data. However, most of these methods do not preserve similarity information, and the information is lost in the training phase. Because similarity information was the main interest in this study, the methods were chosen to preserve similarity information and even to be based on the similarity of data.

On the other hand, similarity information was utilized in diverse, actual industrial applications. The application areas were spot welding and steel manufacturing, including not only low-level manufacturing problems like quality control of spot welding, but also higher-level problems where the aim was to improve the whole manufacturing process. Thus, it was shown that the similarity approach is usable when considering different phases of manufacturing.

In this study, Chapter 4.1 introduced how Bayesian networks and self-organizing maps were used to estimate quality inside a spot welding process. With these methods the training data was divided into classes or nodes based on similarity information. In addition, quality estimates were given based on the most similar events. It was shown that Bayesian networks and self-organizing maps can be used to estimate the quality of individual welding spots and that similarity information is valuable for industrial applications. While with black box models only quality can be estimated, with SOM and Bayesian networks the variable values constituting quality can also be seen, and by avoiding these settings more high-quality products can be manufactured.

Chapter 4.2 also approached similarity information using self-organizing maps. The different states of the spot welding manufacturing process were mapped into a map and an area corresponding to the optimal process state was identified. For new events the most similar node was searched from the map and, based on this information, it could be noted that the ongoing process was drifting from the optimal. In addition, counter actions were performed using the similarity information.

The study carried out in Chapter 4.3 aimed to make the overall manufacturing process more efficient by saving time, money and effort from the beginning of production.
This was done by utilizing information from similar processes when setting up new processes or when searching for suitable quality control methods. In practice, a database of welding events from different manufacturers was built up. The most similar process was sought from the database and its detailed information was retrieved for two distinct purposes: (1) to restore the process parameters leading to high-quality joints and substantially reduce the time needed to set up a new process or (2) to seek the quality prediction and improvement methods already developed for the similar process. In this task also unsimilarity information was found to be valuable in giving information if the most similar event was actually similar enough to be used.

In Chapter 4.4 similarity information was utilized in model maintenance tasks. The study was not done to directly mine any specific information from the industrial point of view, but the aim was to improve overall model efficiency and usability without touching the actual model training procedure. This means the developed methods were not directly modeling the industrial process, but they supported the model modeling the process. In this task similarity information was used in two tasks: to decide the optimal moment for model refitting and to select suitable training data for this fitting. In the optimal-moment-seeking phase, the estimation accuracy of the models for similar data points and the number of similar data points were utilized, while in instance selection the selection was based on a cluster structure where similar data points constituted their own clusters. In both cases similarity information proved its usefulness: an optimal moment for model updating was discovered and the training data set could be reduced to a one-fourth of the original without undermining the model’s accuracy.

As it was shown, research on data mining is spreading into industry. Statistical and machine learning methods can benefit several steps of manufacturing. On the other hand, when working with industry, new purposes for data mining methods can be discovered. The use of real data requires methods that can adapt to changing situations and are flexible and accurate. Thus, real-world applications are also a driving force for the development of new methods and modification of existing ones. This thesis showed some contributions that can be achieved when dealing with an extensive industrial area and demonstrated that industry and data mining research can benefit highly from each other.
References


Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS,
Original publications


Reprinted with permission from IASTED (I), ICSC (II), Springer (III and VI), INSTICC (IV) and IEEE (V and VII).

Original publications are not included in the electronic version of the dissertation.


304. Popov, Alexey (2008) TiO\textsubscript{2} nanoparticles as UV protectors in skin


310. Prokkola, Jarmo (2008) Enhancing the performance of ad hoc networking by lower layer design


316. Marjala, Pauliina (2009) Työhyvinvoinnin kokemukset kertomuksellisina prosesseina-narratiivinen arviointutkimus

Heli Koskimäki

UTILIZING SIMILARITY INFORMATION IN INDUSTRIAL APPLICATIONS

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

ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)