Helsinki University of Technology Publications in Computer and Information Science

Espoo 2003 Report A69

Using visualization, variable selection and feature extraction to learn from industrial data

Sampsa Laine

Dissertation for the degree of Doctor of Technology to be presented with due permission for public examination and debate at Helsinki University of Technology (Espoo, Finland) on the 19th of September, at 12 o'clock noon.

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
FIN-02015 HUT
Finland

Distribution:

Helsinki University of Technology
Laboratory of Computer and Information Science
P.O.Box 5400
FIN-02015 HUT
Tel. +358-50-549 7893
E-mail: Sampsa.Laine@hut.fi

ISBN 951-22-6669-5 (printed version)
ISBN 951-22-6670-9 (electronic version)
ISSN 1456-2243

Otamedia Oy
Espoo 2003

Laine, S., 2003, Using visualization, variable selection and feature extraction to learn from industrial data, Doctoral thesis, Helsinki University of Technology, Espoo, Finland.

Electronic version available in http://lib.hut.fi/Diss/2003/isbn9512266709/

Keywords: Human learning, visualization, variable selection, feature selection, feature extraction, Self-Organizing Map, data mining, statistical analysis

ABSTRACT

Although the engineers of industry have access to process data, they seldom use advanced statistical tools to solve process control problems. Why this reluctance? I believe that the reason is in the history of the development of statistical tools, which were developed in the era of rigorous mathematical modelling, manual computation and small data sets. This created sophisticated tools. The engineers do not understand the requirements of these algorithms related, for example, to pre-processing of data. If algorithms are fed with unsuitable data, or parameterized poorly, they produce unreliable results, which may lead an engineer to turn down statistical analysis in general.

This thesis looks for algorithms that probably do not impress the champions of statistics, but serve process engineers. This thesis advocates three properties in an algorithm: supervised operation, robustness and understandability. Supervised operation allows and requires the user to explicate the goal of the analysis, which allows the algorithm to discover results that are relevant to the user. Robust algorithms allow engineers to analyse raw process data collected from the automation system of the plant. The third aspect is understandability: the user must understand how to parameterize the model, what is the principle of the algorithm, and know how to interpret the results.

The above criteria are justified with the theories of human learning. The basis is the theory of constructivism, which defines learning as construction of mental models. Then I discuss the theories of organisational learning, which show how mental models influence the behaviour of groups of persons. The next level discusses statistical methodologies of data analysis, and binds them to the theories of organisational learning. The last level discusses individual statistical algorithms, and introduces the methodology and the algorithms proposed by this thesis. This methodology uses three types of algorithms: visualization, variable selection and feature extraction. The goal of the proposed methodology is to reliably and understandably provide the user with information that is related to a problem he has defined interesting.

The above methodology is illustrated by an analysis of an industrial case: the concentrator of the Hitura mine. This case illustrates how to define the problem with off-line laboratory data, and how to search the on-line data for solutions. A major advantage of algorithmic study of data is efficiency: the manual approach reported in the early took approximately six man months; the automated approach of this thesis created comparable results in few weeks.

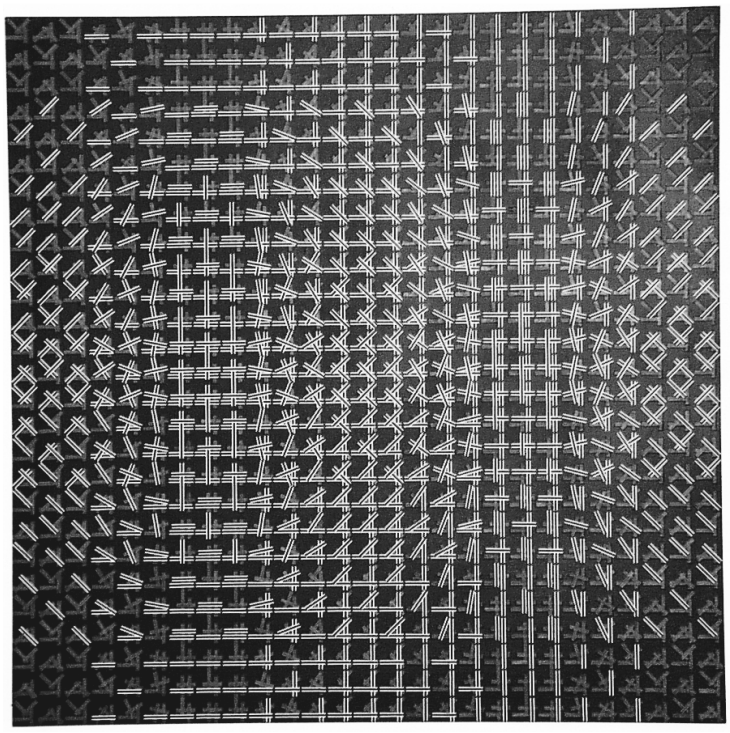
Preface

I'd like to thank the persons that have contributed to this work, directly or indirectly.

Thank you for all of you that have contributed to the scientific results: especially Professor Olli Simula for providing me with the playground, and allowing me to play freely; thank you my opponent, referees and censors, and all persons who critiqued and discussed my thesis. Another significant person is the supervisor of my Licentiate thesis, Professor Sirkka-Liisa Jämsä-Jounela, who started my journey in the enchanting paths of research. Since this thesis aims to support the curious process engineer, I want to mention my role model, M.Sc. Kari Pulkkinen, the metallurgist of the Hitura mine. He embodies the best qualities of a process engineer: craftsmanship, curiosity, and the passion of enhancing the process. In these surroundings, research is a pleasure.

And thank you to all that have contributed to financing this study. In addition of providing the play ground, the laboratory of Computer and Information Science, through Professor Simula, has provided me with great toys to play with. Similar thanks go to Professor Jämsä-Jounela and the Outokumpu Finnmines who resourced my licentiate thesis. These instances allowed me concentrate onto research and pizza, instead of the next rent.

Most importantly, thank you for all persons that bring happiness into my life. My parents and siblings for raising me stable enough to do long-span research. The two choirs I sing with, WiOL and Audite, are paramount: they provide me with friends and music. A happy subconscious allows the conscious brain to do, for example, science.



Would you say that art and science are similar: both try to express intuitions understandably? The above art is by Matti Kujasalo (1994).

List of publications

Publication P1. Laine, S., Lappalainen, H., Jämsä-Jounela, S.-L., On-line determination of ore type using cluster analysis and neural networks, Minerals Engineering, Vol.8 (6), pp. 637-648, June, 1995

Publication P2. Laine S., Ore type based Expert System for Hitura Concentrator, Preprints of the 8th IFAC International Symposium on automation in Mining, Mineral and Metal Processing, Ed. I. J. Barker, Sun City, pp. 321-327, 1995

Publication P3. Laine, S., Pulkkinen, K., Jämsä-Jounela, S.-L., On-line determination of the concentrator feed type at Outokumpu Hitura Mine, Minerals engineering, Vol 13, No 8-9, pp. 881-895, 2000.

Publication P4. Laine, S., Combining Off-Line and On-Line Information in Process Study Using the Self- Organizing Map (SOM), in Proc of IEEE Mountain Workshop of Soft Computing in Industrial Applications, ed. Embrechts J, VanLandingham H, Ovaska S, Blacksburg, USA, pp 71-76, 2001a.

Publication P5. Laine, S., State Based Process Study Using the SOM and a Variable Selection Technique, in Evolving solution with neural networks, the proc. of International Conference on Engineering Applications of Neural Networks, ed. Baratti, R., Caete, J, Cagliari, Italy, p. 15-22, 2001b.

Publication P6. Laine, S., Finding the variables of interest, Minerals Engineering, (15) 167-176, 2002.

Publication P7. Laine, S., Selecting the variables that train a Self-Organizing Map (SOM) which best separates predefined clusters, ICONIP 2002, Singapore, pp. 1961-1966.

Publication P8. Laine, S., Automatic extraction of simple features from process data, ICONIP 2003, Istanbul, pp. 134-137.

Legend of publications

P1 is a comparative analysis between cluster analysis and SOM. The main result was to justify the use of the SOM in data mining. At the time, the SOM was a new and controversial algorithm in the mineral industry.

P2 is an outline of the expert system designed to support operation of the concentrator of the Hitura mine. The main result was an on-line decision support system for process operators. The system comprised an on-line observer of the feed type, and a knowledge base defining proper treatments for each type.

P3 presents the main results of the Hitura work: the expert system that realises the goals defined in P2. These results were based on extensive manual study of the data.

P4 started to automate the manual work that lead to P3. This paper showed how to find the set of on-line variables that contain information of the problem defined using off-line data. In the Hitura case such variables support the creation of the on-line observer of the feed type. P4 was awarded as the best paper in the SMCia/01 (IEEE Mountain Workshop on Soft Computing in Industrial Applications).

P5 continued the work of automating the analysis work required by P3. The paper defined a methodology that shows how to find the process control variables that have the highest leverage in the treatment of a given feed type.

P6 begun to solve the algorithmic problems related to the concepts defined in P4 and P5. This paper proposed a cluster separation measure that allows comparison of spaces of unlike dimensionality, and makes fewer assumptions of the probability distributions of the clusters than parametric measures of cluster separation.

P7 showed, how to find the variables that train a SOM that best illustrate the problem defined as interesting by the user.

P8 proposed a cluster separation measure that allows supervised operation, is robust and understandable. The paper also proposes an approach that uses variable selection to extract features from process data according to the problem definition created by the user. The P8 is the conclusion of the line of research documented in this thesis.

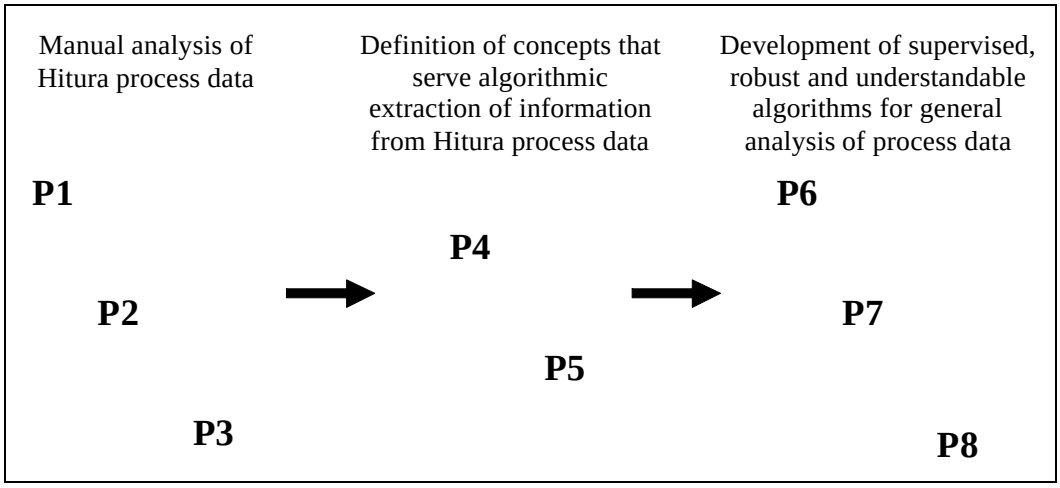


Table of Contents

1 Introduction

The thesis proposes methodologies and tools that help a process engineer to learn from process data. According to my experience most industrial data analysis is based on manual selection of variables and extraction of features. This is time consuming and prone to bias caused by the expectations of the engineer. There is a myriad of statistical theories that support data analysis – why do the engineers not use them? They should, as the data stored in the plant's automation system is a rich collection of full-scale experiments, designed by process operators, geared to solve process control problems.

I believe the origin of the reluctance of the engineers to use statistics is in history. Statistical theories were created by researchers fluent with mathematics, computing with rules and statistical tables. They created tools that demand theoretical skills, but allow sophisticated analyses. However, process engineers do not possess their statistical knowledge: they are bound to insert raw process data into the algorithms, and see if anything useful comes out. Algorithms that are developed to produce accurate results for a given type of data probably produce poor results if fed with raw industrial data. The contribution of this thesis is to propose a methodology, and a set of tools that serves the process engineer in learning from data.

The philosophical discussion is based on the learning theory of constructivism. This theory, proposed in 1970's by Piaget (see e.g. Phillips and Soltis 1998), defines learning as construction of mental models. These models are created to explain and predict the phenomena of interest to a person. The next layer of theory studies organisational learning, based on the concepts of Senge et al (1994). They show how members of organisations learn, and apply their mental models. Then the organisational learning is bound to the statistical theories of learning from data, which allows discussion about how statistics can support organisational learning. Next, the focus is shifted to the properties of the process engineer using the data mining concepts of Glymour (1997). The above theories justify the findings of this thesis: the three criteria designed to assess suitability of a tool for industrial engineers; and a data analysis methodology that complies with these criteria. The goal of this thesis is to support the learning process of the engineer, rather than the learning process of a computer. Elaborated analysis of this philosophy is a main result of this thesis.

The three criteria that benefit industrial learning are supervised operation, robustness and understandability. Supervised (e.g. Jain et al 2000) algorithms allow, and require, the user to define what he[1] considers interesting. They retrieve information related to this problem of interest. Unsupervised algorithms are not guided, but look for the statistically most significant phenomenon. However, results can be statistically significant, but unrelated to the problem of interest. Such results often only confuse the user. Supervised algorithms are more likely to support learning of a given problem. The second viewpoint is robustness; defined by Huber (1981) as insensitivity to small deviations from assumptions. Standard statistical algorithms make strong assumptions of the properties of the data, for example linearity or Gaussian distribution. Use of these algorithms requires the user to pre-treat the data to fit these assumptions, or the

[1] This thesis uses 'he' instead of 'she' or 'he/she' as a matter of practicality. No politics intended.

algorithm may produce poor models. For example, the user may have to remove outliers if he uses Principal Component Analysis (PCA) or manipulate intervariable dependencies to be linear, if he uses linear regression analysis. For industrial engineers pre-treatment is a problem, as they often have neither the statistical skill nor the time to modify the data to serve the algorithm. The only pre-treatment required by the algorithms of this thesis is the scaling of variables, e.g. to unit variance. The third criterion is understandability. Statistical experts can understand and skilfully use the complex algorithms of literature; process engineers may not understand the principle of analysis, parameterization, or the output of such algorithms. This thesis advocates simple algorithms that the engineers understand: according to Glymour (1997) users prefer understandability of algorithms over their accuracy. Study of the above three criteria is another main result of this thesis.

This thesis proposes a data analysis methodology that complies with the above criteria. It is based on three statistical services: visualization, variable selection and feature extraction. Visualization is paramount in any learning from data; consider e.g. the scatter plot, which is a standard tool in most statistical softwares. Variable selection and feature extraction are necessary when browsing the hundreds of signals stored by the process automation system. Supervised variable selection can look for variables that contain information relevant to the user; supervised feature extraction can create simple models that clarify the illustration of the phenomenon of interest, and allow the user to study the models to learn how to read his variables. This methodology geared to support practical analysis of industrial data is the third main result of this thesis.

The above theoretical and practical concepts are demonstrated by the case of the Hitura mine. It shows how the approach of this paper supports construction of mental models in the mind of the operator. The early work reported in publications P1-P3 demonstrates how to manually select variables and extract features. The publications P4 and P5 describe a framework for algorithmic analysis; and the publications P6-P8 propose algorithms that realise this framework. The conclusions of this work are presented in this report. The developed algorithmic approach has to advantages over the manual approach: objectivity and efficiency. Objectivity is demonstrated indirectly: the manual study (publications P1-P3) provided results that we knew to look for – the goal was to prove already suspected mechanisms; the algorithmic analysis provided unexpected results that inspired learning. Efficiency is demonstrated by the Hitura case: manual study required approximately six man months; algorithmic study created comparable results in less than a week. The tools proposed in this thesis can also be used to analyse other data, as shown by the analyses of data from public domain data repositories (publication P8) and data of fuel emissions (Laine 2002).

This thesis is organised as follows. The chapter 2 discusses the theories of learning and data analysis; the chapters 3, 4 and 5 study the algorithms of visualization, variable selection and feature extraction. The chapter 6 uses the Hitura case to illustrate the approach of this thesis. Chapter 7 concludes this thesis. A reader with limited time may only want to read the case study of chapter 6 and the conclusions of chapter 7.

2 Learning from process data

Learning has been studied for a long time: already Plato created theories of learning (see e.g. Phillips and Soltis 1998). This chapter starts from abstract learning theories, and progresses towards practical aspects of data analysis. The goal is to study how statistics can serve engineers.

2.1 Theories of learning

The oldest learning theory considered here is behaviourism, created by Watson in the early 20th century. According e.g. to Phillips and Soltis (1998) it considers the learner to learn automatic responses to stimuli, like the dogs of Pavlov that learned to salivate when a bell was rang. The mental processes of learners are considered very simple.

This thesis is based on the more complex learning theory, constructivism, which describes learners as active players that collect information and construct mental models to model the environment (Phillips and Soltis 1998). Every learner is exposed to a unique chain of experiences, and thus constructs unique mental models. A teacher cannot and should not try to unify the models, but support the student's personal perception of reality, and creation of personal mental models. This is the philosophical basis of this thesis: not to create complex computer models that create 'the right answers', but to provide the user with information related to the problem of interest, and let the user construct the models into his mind.

The next level of theory is organisational learning, which describes how individual mental models influence the behaviour of organisations. According to Wheatley (1991), a modern organisation is a web, where information and resources flow from person to another. Employees, the knots of this web, must continuously learn, as the web keeps changing. The employees must, and have the opportunity to, learn continuously. Such learning is also the basis of the theories of Senge et al's (1994). This thesis describes two of their theories that are related to the construction and use of mental models: the wheel of learning and the ladder of inference.

The wheel of learning is presented in Figure 1. The learner alternates the phases of hypotheses testing and hypotheses creation. A learner observes a process (observing), and collects data that he finds interesting. He compares this data against his mental models (reflecting): if there are discrepancies, he constructs new mental models to explain the measured phenomenon. The he probably wants to test these hypothetical mental models, and plans a series of tests (deciding). In the fourth phase (doing) he deploys this plan, and starts a new round of the wheel. This is the common approach of trial and error.

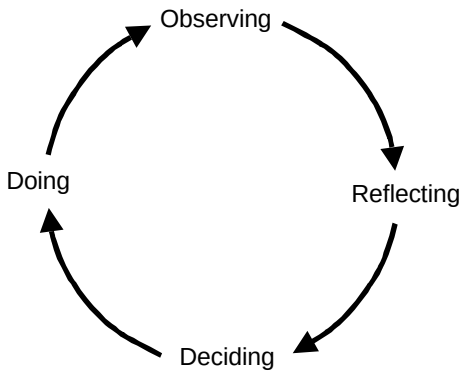


Figure 1. The wheel of learning (Senge et al 1994)

The ladder of inference is presented in Figure 2. While the wheel of learning describes the overall process of learning, the ladder zooms into the phases of observing, reflecting and deciding. The first step of the ladder is the selection of data the user considers relevant. Note, that persons with unlike mental models probably select different data. A person with a doctoral thesis on temperature measurement is bound to study temperatures. At the next step of the ladder, the user adds meanings to the collected data: for example, person 1 may consider the value 210 to be high; person 2 may know that, for tank B, this value is nominal. Then the person makes assumptions to fit the collected data as input to an existing mental model. The person may e.g. disregard one of the collected values if it is abnormal in respect to the other data. Then, these data are input to the model that best fits the input; and the model produces conclusions. If the process of inference was run smoothly, that is, the data matched the used mental model well, the user accepts the conclusions with high certainty, and adopts them as beliefs. According to Senge et al., the above process takes only seconds, and is run on the subconscious level.

According to Senge et al. this subconscious inference may cause misunderstandings and heated discussions. Consider a scientific debate where both scientists have access to the same data. With their unlike mental models, they are likely to select different subsets of the data, add different meanings, make different assumptions, draw different conclusions, and adopt different beliefs. If these beliefs are adopted with high certainties, and they are dissimilar, there may be a conflict, as both persons feel their conclusions are based on objective data, and are correct. The remedy, according to Senge et al., is to slow down and justify the ascension of the ladder. This is in accordance with the philosophy of science: to state how the scientist drew his conclusions. When practiced in organisations, explicated ascension of the ladder ideally results in appreciation of the other person’s views, and alignment of the mental models of the group of experts. Optimally, this leads to the unification of the efforts of the group. The tools of this thesis aim to help the engineers to explicate their ascension of the ladder of inference.

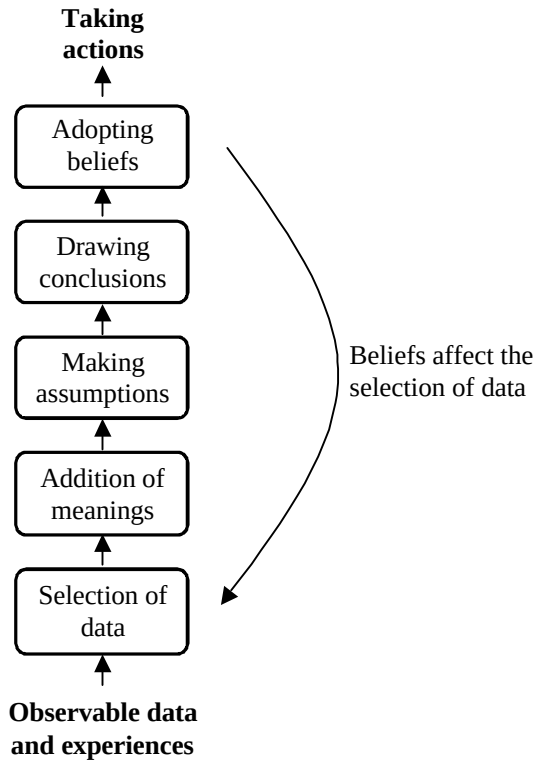


Figure 2. The ladder of inference (Senge et al 1994)

The research of the learning organisation has been critiqued by Easterby-Smith (1997). The main critique is reliance on any single theory, for example, on the above theories of Senge. Easterby-Smith suggests that six areas of science should be considered when defining the learning organisation: psychology, management science, sociology and organizational theory, strategy, production management, and cultural anthropology. This thesis attempts to follow this lead by binding the theories of Senge into the theories of statistics and data mining, and eventually, to the reality of the Hitura concentrator.

2.2 Analysis processes for learning from data

This section advances the discussion of this chapter towards practical data analysis by looking at the theory of data mining. According to Fayyad et al (1996b), data mining is a window between statistical sciences and computer sciences. These algorithms are intended for the study of large data bodies, for example, of industrial data. Data mining techniques have been surveyed e.g. by Chen et al (1996). This section describes two data mining models: the CRISP-model and the KDD-process.

The CRISP model of Chapman et al (2000) is presented in Figure 3. The CRISP-cycle starts by business understanding: to learn the business domain and to define the goals of analysis. According to Hand (1994) statisticians often ignore this phase, and rush to apply their favourite algorithm. According to Pyle (1999), business understanding determines 80% of an application's chances of success. The peer of business understanding is data understanding: an assessment of the potential and reliability of

the data. These two tasks can be supported by the tools of exploratory data analysis, see e.g. the surveys of Lowe and Tipping 1996 and Vesanto (2002). Together, business and data understanding produce a data analysis plan, which defines, how to attempt solving the pending business problem. The next pair of phases prepares data for modelling, and calculates the model. Data preparation is especially important in the analysis of industrial data, as the artefacts in the data, created e.g. by process or instrumentation disturbances, may disturb the selected modelling approach. The result of this pair of phases is a model that aims to respond to the question defined in the first pair of phases. The next CRISP-phase is evaluation, where the user assesses whether the answer is reliable, and responds to the question. If not, the CRISP-loop is re-entered to create enhanced questions and answers. When satisfied, the analyst deploys the results to solve the pending problem. Note the similarity of this iterative CRISP-model to the wheel of learning (Figure 1).

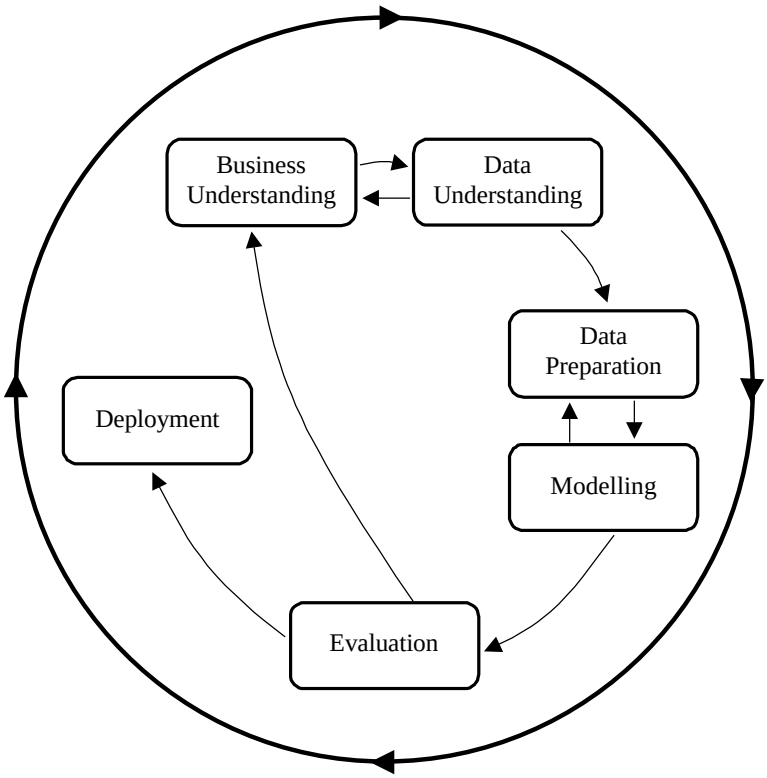


Figure 3. The Crisp Data Mining model (Chapman et al. 2000)

The KDD-process, created by Fayyad et al. (1996b), is shown in Figure 4. The KDD-process can be considered as a zoomed view of two CRISP-phases: data preparation and modelling. The KDD-process starts by the selection of the variables and data points that are considered relevant for the study at hand. In the second KDD-step, the user pre-processes the data: e.g. removes outliers, fills in missing values of data, and scales the variables (see e.g. Hätönen et al 2003). The pre-processed data is transformed to further prepare it for the selected modelling algorithm: for example the dimensionality of data can be reduced to facilitate computation, or features can be extracted to expose the relevant information. The model is identified with this data with the goal of capturing the essence of the studied phenomenon. If the model is

assessed as reliable, the user can study the model to learn of the studied process. Note the similarity of this KDD-process to the ladder of inference (Figure 2).

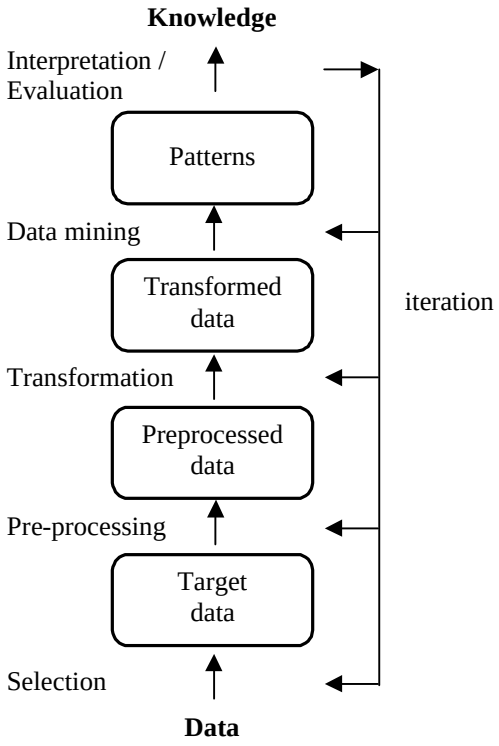


Figure 4. The process of Knowledge Discovery from Databases (KDD) (Fayyad et al. 1996b)

The KDD-process also resembles the ladder of inference in the sense that subconscious mental models influence the results. In the first step of both models, the user selects data he considers relevant, pre-processes and transforms it according to requirements of the used modelling approach, and often accepts the results as the truth, although several subjective steps have lead to the results. The goal of this thesis is to create a data analysis methodology that supports the engineers to explain and justify their choices: it supports the user in the definition of the problem, selection of data, extraction of features, and provides understandable results whose reliability can intuitively be assessed. Justification of this methodology is the essence of this thesis.

2.3 Areas and preferred properties of the statistics of data mining

Algorithms that serve the above data analysis methodologies have been proposed by hundreds of scientists. The next three chapters study individual algorithms; this section maintains a more general level by discussing the areas of statistics relevant to data mining, and desirable properties of such algorithms.

2.3.1 Areas of statistics relevant to data mining

The Table 1 presents the areas of statistics that, according to Glymour et al (1997), are relevant to data mining.

Table 1. Areas of statistics according to Glymour et al (1997)

Area of Statistics	Description of activities
Exploratory data analysis	Hypotheses generation rather than hypotheses testing
Experimental design and sampling	How to select cases if one has liberty to choose
Statistical graphics	Data visualization
Statistical modelling	Regression and classification techniques
Statistical inference	Estimation and prediction techniques

Exploratory data analysis is useful when browsing unfamiliar data. These tools help to create useful hypotheses, that is, to ask meaningful questions. Such questions are formulated in the reflection and decision phases of the wheel of learning. For further discussion of exploratory data analysis, see Lowe and Tipping (1996) or Vesanto (2002). Vesanto, for example, discusses the use of the Self-Organizing Map (SOM) in exploratory data analysis. The problem definition phase of the methodology of this thesis is based on exploratory data analysis.

Experimental design allows the user to design test runs that help analysing a given hypothesis. For example, experimental design can show how to manipulate the process parameters to create useful data. Sampling theories then help to select the variables and data points that serve analysis of this hypothesis. Experimental design and sampling supports the deciding and observing phases of the wheel of learning. The variable selection tools of this thesis can support the task of sampling.

Statistical graphics is an important part of most data analysis tool-kits: for example, the scatter-plot can be found on most data analysis softwares. Visualization of data supports the reflection phase of the wheel of learning as it helps the study of the phenomenon of interest. Discussion of visualization tools is elaborated in the chapter 3.

Also statistical modelling supports reflection. For example, linear regression (e.g. Kachigan 1991), may reveal dependencies between variables. Also feature extraction, discussed in chapter 5, performs statistical modelling with the attempt of extracting relevant information from data. The user can study the created models to learn how the phenomenon of interest manifests in the data.

Statistical inference is similar to statistical modelling; but it operates on a more abstract level. A central concept of statistical inference is probability distribution. For example, the algorithms of Bayesian analyses perform statistical inference. These algorithms can be used, and their use develops a thorough understanding of the studied domain. Such knowledge allows e.g. explicit assessment of the reliability of results.

2.3.2 Preferred properties of statistics of data mining

Glymour et al (1997) prefer three properties of data mining methodologies: clarity about goals, appropriate reliability assessment, and adequate accounting for sources of uncertainty.

Clarity about goals is a basic requirement. While the definition of this goal is a task of the user, he should be supported in his task. If this fails, the user may get correct answers, but to a wrong question (Hand 1994). A good methodology supports the user in a clear definition of the analysis goal. Recall the CRISP-loop; the clear definition of the goals is represented by the first pair of phases: problem and data understanding.

Appropriate reliability assessment is related to understanding whether the algorithm at hand can serve the analysis of the given data; this is close to understanding the strengths and limitations of the algorithms. For example, is the size of the data suitable for the algorithm, or does the data fit the assumptions of the algorithm. This task is realised in the second pair of tasks of the CRISP-loop: the user needs to pre-treat the data to serve an algorithm that can answer the questions of the user.

Adequate accounting for sources of uncertainty is related to the evaluation of the result, and to the assessment of error sources. Good algorithms support the user or algorithm in analysing, whether the results are reliable. Bayesian techniques provide an explicit analysis of such uncertainties. They have been used, for example, to train Artificial Neural Networks (Lappalainen and Honkela 2000, Lampinen and Vehtari 2001) and spline-based regression analysis (Smith and Kohn 1996). Considering the CRISP-loop, this is the phase of evaluating the usefulness of the result.

2.4 The methodology of this thesis

The themes of the previous section are reviewed from the viewpoint of this thesis: the requirements of a process engineer analysing industrial data. The core of this thesis is presented in Figure 5: it shows the tools that should belong to an engineer's toolkit, and the properties that these tools should have. This section elaborates the three criteria of Figure 5, propose a methodology that serves the process engineer, and defines the three tools that are required to realise the methodology.

2.4.1 Preferred properties of statistics of this thesis

This thesis advocates three properties: supervised operation, robustness and understandability.

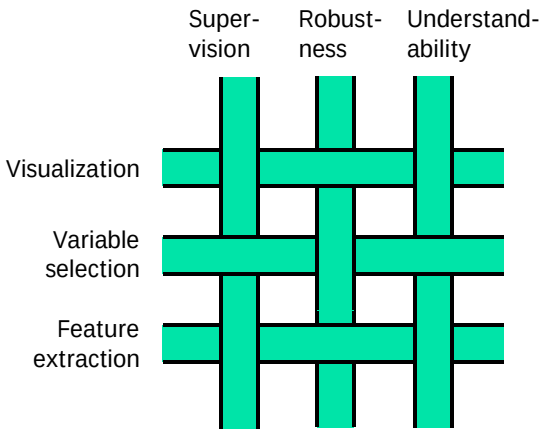


Figure 5. The conceptual basis of the tools created for this thesis

Supervised algorithms allow, and require, the user to define what he considers interesting. This is close to the first theme of Glymour et al discussed above, supporting a clear definition of the goal of the analysis. Supervised tools can use this definition to retrieve, analyse and visualise relevant data. A simple supervised algorithm is linear regression, where the user defines the problem by pointing out a variable that he wants to study. Algorithms can then automatically retrieve variables and identify parameters that create the most accurate model to explain the variance of this variable. Another supervised algorithm is the classifier, which selects the variables and parameters that define a model, which best reproduces the labels of the training data. The user can study these models of linear regression and classification to learn about the phenomenon defined interesting by selecting the variable or labelling the data. However, if the user does not know, what is of interest, he must use the tools of exploratory data analysis to learn enough to define the first hypothesis of the problem. This definition is an integral part of the methodology described in the next section.

Robustness is close to the second theme of Glymour et al: assessment of whether an algorithm can serve the task at hand. Robust algorithms are versatile: they can analyse a variety of data. The more assumptions the algorithm makes, the less likely it is, that the available data can be analysed with it without elaborate pre-treatment of the data. For example, if the algorithm assumes Gaussian distribution, it is unrobust towards outliers. Industrial data is polluted by outliers created by process and instrumentation disturbances. In industrial surroundings, the user must, usually manually, clear the data of outliers before applying the algorithm. Another assumption leading to unrobustness is assumption of linear dependencies between variables. If the user uses e.g. correlation analysis, he may have to pre-treat the data e.g. with non-linear mappings to linearise the dependencies, before he can achieve accurate linear models. Scientific researchers have time and statistical knowledge to clean the data of elements that do not fit the used algorithm; industrial engineers do not. They probably insert raw industrial data into the algorithm and study the results. If used in such a way, unrobust algorithms are unreliable. This thesis advocates robust methods, as they allow the analysis of uncleaned industrial data.

The most important of the three themes may be understandability. According to Glymour et al. users are often more interested in understandability than accuracy. Understandability is related to the second and to the third theme of Glymour et al. If the user understands the principles of the algorithm, he can assess whether it can serve

the task at hand. And if in addition, the parameterisation and the output are understandable, the user can assess whether the results reliably answer his problem, and where are the possible sources of uncertainty. Consider a common engineer, or a good scientist: they will publicly advocate only the results they can understand and justify.

There are some statistical themes that have been excluded from the scope of this thesis. Possibly the most debatable is the exclusion of probabilistic analysis and Bayesian inference. This thesis acknowledges their theoretical benefits, and the explicit analysis of the reliability of the result. However, this thesis argues that they are not understandable to the process engineer. I consider the definition of a Bayesian model to be comparable to the definition of an expert system. According to Bearman and Milne (1992), elicitation the knowledge from process operators and encoding it into the models of the expert system may be too difficult a task to be economically justifiable. For example, Gaag et al (2002) found definition of the probabilities required by the Bayesian approach difficult. This next section describes a simpler approach that aims to construct models into the minds of the experts, instead of coding mathematics to computers.

2.4.2 The analysis process of this thesis

Figure 6 illustrates the analysis process proposed by this thesis. The main function of this process is to support the user in defining the problem. The rest of the scheme is straightforward application of the algorithms of variable selection, feature extraction and visualization, described in the next subsection. As all learning schemes presented previously in this chapter, also the proposed approach is iterative: the user probably evolves in his ability to define relevant questions.

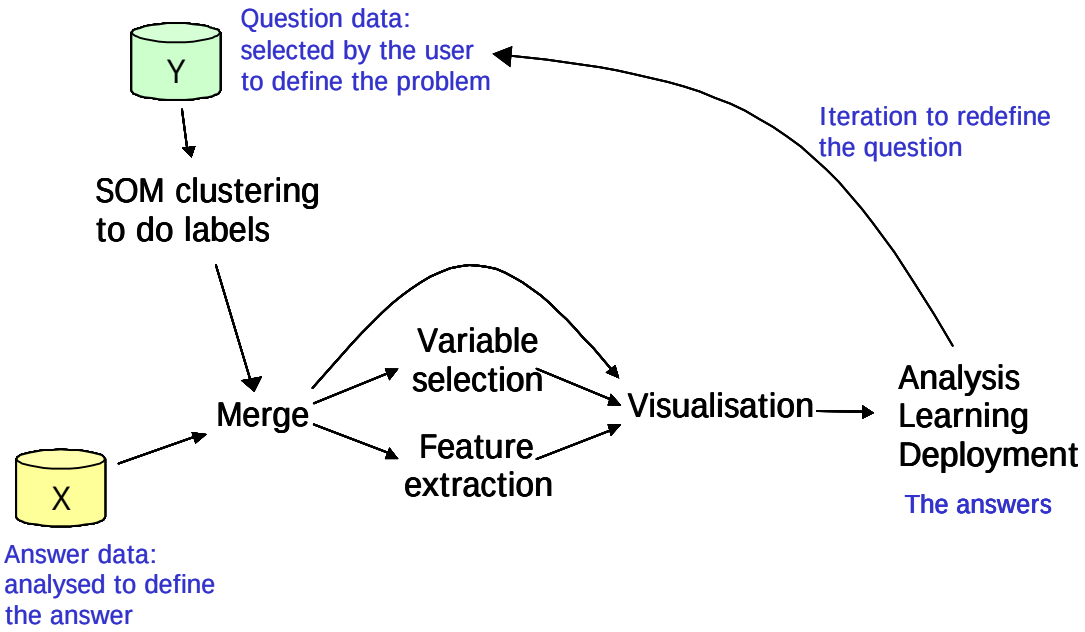


Figure 6. Use of Y-data to label the points of X-data

The user starts the analysis by selecting a set of Y-data that is suspected to contain information of the problem of interest. In the Hitura case (chapter 6) the problem is volatility of the process feed type: the variables of the Y-data comprised off-line laboratory analyses of the mineralogy of the process feed. The on-site laboratory of Hitura has developed into an accurate and versatile source of process control information. In my experience, most industrial plants have such an on-site laboratory. In addition of selecting the Y-variables, the user selects the time-periods he wishes to study, for example, one period that manifests the interesting problem, and one period that manifests nominal conditions.

The user studies this data, for example, with the Self-Organizing Map (SOM). The SOM is useful as it provides a topologically ordered mapping of the typical patterns of the data, in the Hitura case, the types of process feed. If these patterns match with the user's mental model of the problem, the definition of the problem has been completed. If the SOM does not describe the user's problem, he iterates his selection of the Y-data, his understanding of the problem, or concludes that the data does not allow definition of the problem. Another benefit of the SOM is that it supports joint work of a group of experts: an expert can justify his statement by referring to the patters visualised by the SOM. The possibly emerging shared understanding of the problem is a sound basis for further analysis. Such a user-driven approach is supported by Phillips and Soltis (1998), who claim that a person can only learn of issues that he already knows something about.

The analysis continues by using the SOM to create the labels required by the supervised tools of this thesis. They are created by clustering the SOM according to the method of Vesanto and Alhoniemi (2000). This process is described in subsection 3.2.2. Table 2 illustrates a synthetic example, where the SOM, trained with Y-data, is used to define two feed types *serpentinite* and *amphibole* for a comparative study. The unlabelled rows are ignored in the analysis. The user needs to select at least two feed types, as the method of this paper is based on studying the separability of clusters defined by these labels. The upper limit to the number of clusters is, in practice, defined by the understandability of the results; and in theory by keeping the clusters large enough to ensure statistical significance. To conclude, the definition of these labels concludes the first phase of the methodology – the definition of the problem.

Table 2. A synthetic data set of labelled data

	Y_1	Y_2	*Label*	X_1	X_2	…	X_n
Day 1	3.2	65.4	serpentinite	-4.4	-4.4	…	-4.4
Day 2	2.5	45.8		-3.6	-3.6	…	-3.6
Day 3	1.6	24.4		1.2	1.2	…	1.2
Day 4	2.3	22.5	amphibole	2.3	2.3	…	2.3
Day 5	1.4	23.3	amphibole	4.3	4.3	…	4.3
…	…	…	…	…	…	…	…
Day N	5.2	64.4	serpentinite	-5.4	-5.4	…	-5.4

The second phase of this methodology studies another body of data, X-data, to define an answer to the question of the user. The first step is to merge the labels into the X-data. The user must define the link between the points of the Y- and X-data. A straightforward link is the timestamp, illustrated in Table 2: both data sets are day

average measurements of the same process. After establishing this link, the labels are attached to the X-data, and the values of Y-data are discarded from further analysis.

The basic tool of this second phase is variable selection. It supports exploratory data analysis, as it finds variables that contain relevant information; these variables may have been unknown to the user. For the same reason, variable selection supports experimental design and sampling. Certain variable selection algorithms e.g. the approach of this thesis, support dimension reduction, which in turn supports statistical modelling and inference (Back and Trappenberg 2001). In the study of industrial data, a tool of variable selection is especially useful, as the user must select the relevant variables from the hundreds of variables stored in the plant's automation system. Chapter 4 discusses algorithms of variable selection, and introduces the algorithm proposed by this thesis.

An auxiliary tool of the second phase is feature extraction. It supports the same tasks as variable selection, but offers extended service: it creates mathematical models that aim to enhance the visibility of the studied problem in the data. The original variables may not themselves describe the problem. A simple example is to use the derivative of the signal instead of its absolute values. The feature extraction approach of this thesis is geared to support learning, instead of enhancing the accuracy of the models. The user can study the found feature extraction models and enhance his abilities to read his data. For example, learns to look at the trend of the data instead of its absolute values. The chapter 5 discusses the algorithms of feature extraction, and introduces the algorithm proposed in this thesis.

The benefit and the drawback of the above approach is its emphasis of problem definition. If the user can provide one, the analysis is greatly supported. If not, the method is not applicable. However, the first phase of the methodology is geared to support the user's definition of the problem. The tools of exploratory data analysis (e.g. Lowe and Tipping 1996) serve this work. According to Vesanto (2002) the SOM serves exploratory data analysis well. This is the main justification for the selection of the SOM as one of the bases of the proposed analysis process.

From the viewpoint of practical process control, the benefit of this methodology is emphasis of construction of mental models in the minds of the analysts, instead of creation of mathematical models into computers. Mathematical models are, in industrial surroundings, difficult to maintain: as the process flow sheet, raw materials, or product specifications change, the model must be redefined. If this redefinition is laborious, conceptually complex, or if the application is not crucial, this work may be neglected, and the computers start to produce biased results – the application loses credibility. The mental models of the engineers and process operators are more flexible: they are the persons that daily consider the changes in the evolving process. It is more practical to support their learning than to continuously model their expertise.

Finally, recall the ladder of inference (Figure 2, page 5). The above methodology is geared to understandably provide information that is related to a problem of interest. This allows the members of a group of experts to refer to data when they want to justify their ascension of their ladder of inference. A person can define which process phenomena he is discussing, what is the significance of the visualised patterns of data, and what his conclusions are. Ideally, such justified discussion leads to shared understanding of the problems, and to coherent operation of this team of experts.

2.5 Summary and discussion

As globalisation advances, markets become more competitive. Engineers of industry must adapt to new technical and economical configurations: constantly learn new skills. This chapter started by studying basic learning theories, and their application to analyse learning organisations. The objective of this thesis is to provide a foundation for the tool of this thesis, which are geared to support learning from process data.

The field of statistics has created a plethora of approaches that serve analysis of data. This chapter presented two of them: the CRISP-model and the KDD-process; and bound these statistical theories to the psychological and organisational learning theories. Then this chapter discussed the fields of statistics that support data mining, and the desirable properties of these algorithms. This thesis claims that advanced statistical algorithms do not serve the average industrial engineer; only few engineers use them.

The goal of this thesis is to propose a methodology that can be accepted by the engineer. This development work is based on three criteria: supervised operation, robustness and understandability. The user supervised the analysis by defining the problem with a set of data known to contain relevant information. This thesis proposes the clustered SOM to be the basis of this work. Then the supervised algorithms of variable selection and feature extraction are used to study the data to find information that is related to the problem defined interesting by the user. This information is visualised to the user.

A major contribution of this thesis is to emphasise and support the process of problem definition. And to create tools that are explicitly guided by this definition. The three tools are introduced in the next three chapters; and the process of supervised data analysis in chapter 6 by the analysis of the data of the Hitura concentrator.

3 Visualization

Visualization is a major asset when learning from data (Keim et al 1996, Keim 2002). It supports the reflection phase of the wheel of learning, as it supports the user's task of comparing the behaviour of the process against the user's hypotheses.

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

This chapter starts by surveying the algorithms of literature using the three criteria of this thesis; and then discusses the visualization algorithms chosen for this thesis.

3.1 Survey of visualization algorithms

According to Keim and Kriegel (1996), there are five visualization algorithm types that serve the study of large data bases: pixel-oriented, geometric, icon-based, hierarchical, and graph-based techniques. For another survey, see Tufte (1983, 1990). In pixel-oriented techniques each data point is allocated a pixel in the visualization. This allows large amounts of data to be represented in one figure. In geometric techniques the data is mapped onto a surface that supports visualization. In icon-based techniques each data point is allocated an icon that describes the data in the point; a well-known example of this are Chernoff faces (Chernoff 1973). In hierarchical techniques the variables are arranged into a hierarchical set of e.g. two-dimensional subspaces that allow scatter-plot based visualization. In graph-based techniques the task is to visualize a graph as understandably as possible. This section discusses some of these algorithms using the criteria of supervision, robustness and understandability.

The difference between unsupervised and supervised visualization is illustrated in Figure 7. In the unsupervised plot, the user can only study the general properties of the data, for example, correlations or a cluster structure. The left subplot would not invite the user to further study these two variables. The data in the right subplot was augmented with labels that encode the user's interests – in this case the mineralogy of the process feed type. This plot shows that the two variables have mutual information; a conclusion that is not enforced by the left subplot. The label-augmented scatter plot is the main visualization tool of this thesis.

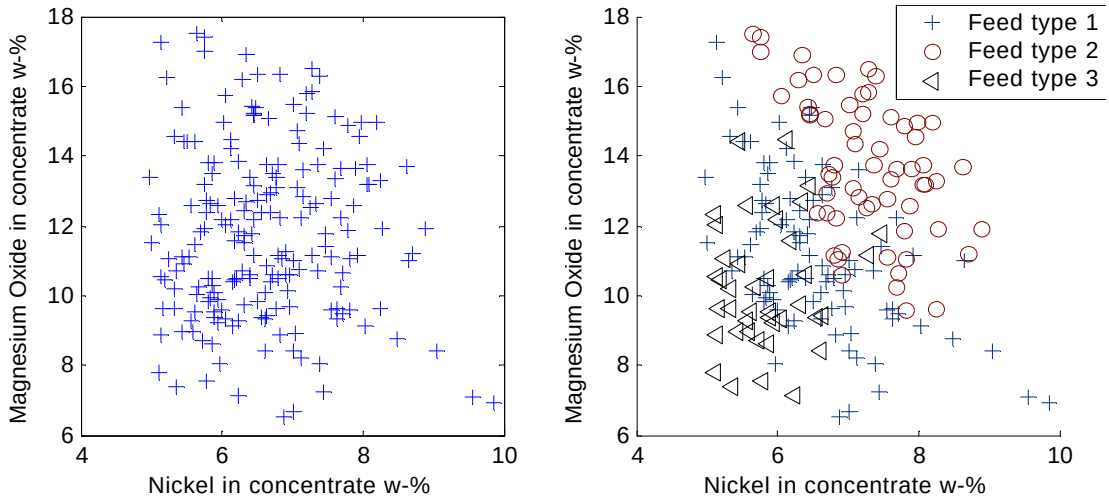


Figure 7. Unsupervised (left) and supervised (right) visualization of two variables

Sometimes more than two variables are required to describe a phenomenon; the user must use a dimension reduction algorithm, as scatter-plots can conveniently illustrate at most three-dimensional spaces. Unsupervised and supervised dimension reduction is illustrated in Figure 8. The Principal Component Analysis (PCA) (see e.g. Kachigan 1991) is described by the solid line. This unsupervised method aligns its principal components to capture as much of the variance of the data as possible. In this sense, the PCA illustrates the dominant phenomena in the data. However, the two clusters are mapped on top of each other: their differences are not illustrated. Supervised dimension reduction, e.g. Fisher Discriminant Analysis (FDA) (Fisher 1936), described by the dashed line, is better equipped to describe the differences of the clusters, as its identification uses the label information (clu_1 or clu_2). The FDA spans a space in which these clusters, modelled as Gaussian ellipsoids, separate as well as possible.

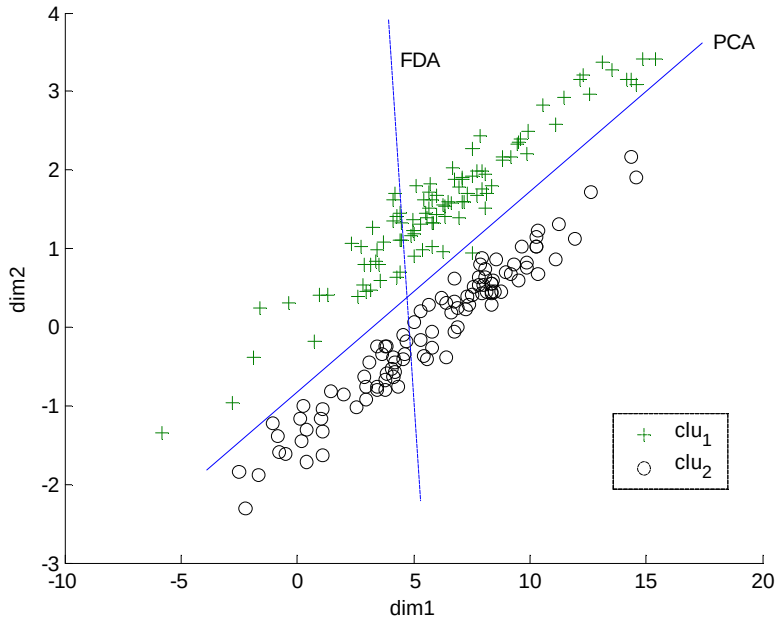


Figure 8. The first components returned by PCA and FDA. The unsupervised PCA aligns the component along highest variance in data; the supervised FDA aims to separate the clusters.

Robustness is discussed by studying the influence of an outlier, and of linearity. Most industrial data contains outliers, produced e.g. by process or instrumentation disturbances. The used tool should be robust against them. For example, a common tool, PCA, is not robust, as it models the clusters with the variance-covariance matrix. In the calculation of variance, see Equation 1, a data point gains weight according to the square of its distance from the main body of data. As illustrated in Figure 9, a single outlier can bias the analysis. Following this reasoning, all methods based on squared distances are more vulnerable to outliers than methods based on linear distances. Such unrobustness influences, for example, linear regression and correlation analyses.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1)$$

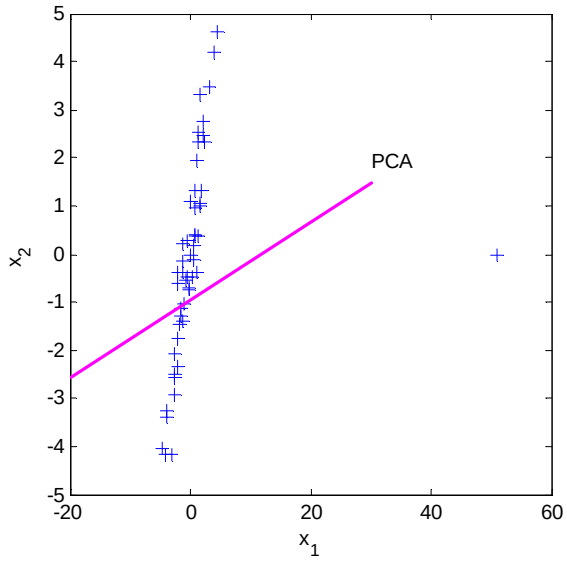


Figure 9. The Principal Component Analysis (PCA) result biased by an outlier

A number of common dimension reduction algorithms, such as the PCA and FDA are based on linear mapping of the data. Its basic problem is illustrated in Figure 10: the linear PCA model (solid line) does not well describe the non-linearly distributed clusters. The PCA projection would map three of the four clusters on top of each other. The non-linear SOM (dashed line) allows a more faithful analysis of the clusters. The non-linear approach is more flexible in the analysis of data.

The third criterion is understandability. The common scatter-plot is the favourite of this thesis. If spaces with more than two dimensions need to be studied, e.g. for the definition of the problem, this thesis prefers the SOM. According to Vesanto (2002), and as discussed below, the SOM is a good tool for exploring high-dimensional data. However, the SOM is an unsupervised algorithm. While this does not handicap problem definition, it handicaps visualization of the results, as illustrated in Figure 7. If high-dimensional spaces need to be studied in a supervised manner, the user can use e.g. the non-Gaussian FDA proposed by Torkkola and Campbell (2000), or the supervised SOM of Kaski et al (2001) and Kaski (2002). However, this thesis mostly prefers to request only two-dimensional spaces from the algorithms of variable

selection and feature extraction, as their results can be illustrated with the common scatter plot.

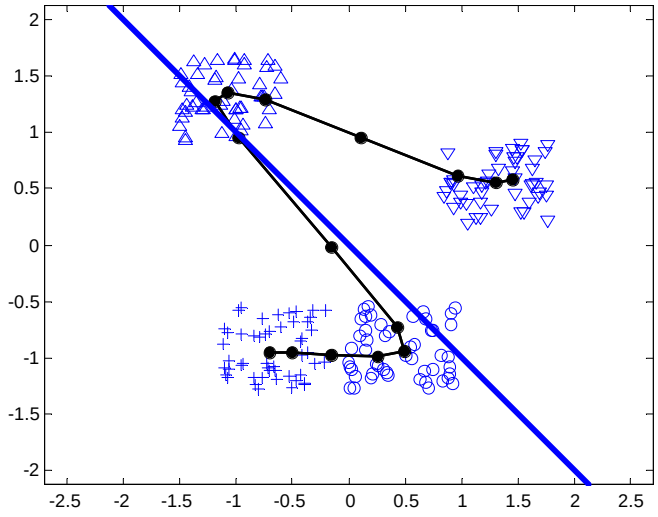


Figure 10. Illustration of the capability of two dimension reduction algorithms to project clustered data into one dimension. The 1-dimensional SOM (curved dashed line) is better equipped to describe the clusters than the first PCA component (linear solid line).

3.2 The approach of this thesis: scatter-plots and the SOM

This thesis uses scatter-plots and the SOM for visualization. Although the scatter-plots are the preferred visualization algorithm of this thesis, related theory is bypassed, as the scatter-plot is a well-known tool; discussion of visualization tools is elaborated e.g. by Tufte (1983, 1990), Keim and Kriegel (1996), and Keim (2002). This section studies the SOM. Note that the use of the SOM is not obligatory in the methodology described in the subsection 2.4.2; the user can define his problem with any tool that provides the labels for the data points. This thesis uses the SOM because, according to Kohonen (2001), it is a well established tool for visualising high-dimensional data.

3.2.1 Training the SOM

The SOM, created by Kohonen (e.g. Kohonen 2001), is an unsupervised Artificial Neural Network (ANN). An ANN can be described as a graph with neurons as nodes and connections as arcs. The SOM used in thesis can be considered a two-dimensional graph, like the honeycomb of a beehive (see e.g. Figure 12); the nodes are the units of the honeycomb, and arcs connect each unit to its immediate neighbours.

In training, this graph is stretched through the densely populated areas of the space, as illustrated in Figure 11. The vertices try to reach dense areas of data, and the stretching

edges oppose this process. These edges keep neighbouring vertices close each other in the training data space. The attraction of the vertices towards the densely populated areas allows the SOM to be used for vector quantization (Györfi 2002). The edges allow the topology of the data to be preserved: data points close to each other in the training data space are mapped close to each other on the SOM, which allows the to function as a dimension reduction technique. This property resembles Sammon mapping (Sammon 1969). The combination of vector quantization and dimension reduction allows the SOM to visualise large bodies of high-dimensional data.

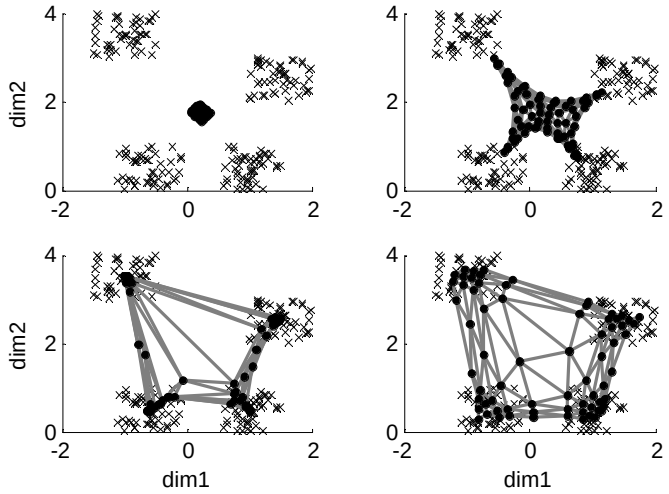


Figure 11. Training the SOM grid to span over the data of four clusters. The first subplot in top-left shows the initialised SOM; the next three subplots show the SOM after 1, 2 and 4 (=final) epochs of training

Mathematically, the position of each neuron m_i in the N-dimensional space is defined by $m_i = [m_{i\ 0}\ m_{i\ 1}\ \ldots\ m_{i\ N-1}]^T$. The first step of training is to initialise the m_i e.g. randomly or by arranging the nodes along the plane defined by the two first principal components of the training data. In training, data points are sequentially introduced to the SOM. In each introduction, the SOM neuron $m_c(t)$ closest to the data point $x(t)$ is selected with Equation 2. This neuron is the Best Matching Unit (BMU) (Kohonen 2001), sometimes also called the winner.

$$\forall i, \|x(t) - m_c(t)\| \leq \|x(t) - m_i(t)\| \tag{2}$$

The winner and its neighbourhood are moved towards the introduced data point according to Equation 3; the neighbourhood is defined by $h_{c(x),i}$ This thesis uses the Gaussian neighbourhood illustrated in Figure 12.

$$m_i(t+1) = m_i(t) + h_{c(x),i}\big(x(t) - m_i(t)\big) \tag{3}$$

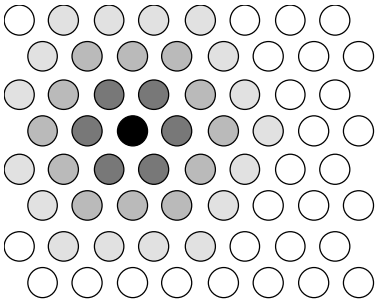


Figure 12. The winner neuron (the black circle in the centre of the shaded area) with its Gaussian neighbourhood (the shade indicates the membership in the neighbourhood)

In the early stages of training, the radius of the neighbourhood is large; and most of the SOM neurons strongly belong to any node's neighbourhood. This creates good global ordering of the SOM visualization. As the training progresses, the radius is reduced to allow good local ordering. Figure 11 illustrates this process: in the early training stages the SOM learns to roughly cover the space; in the late stages the SOM fine tunes to describe the local details.

The trained SOM approximates the probability density function of the training data; at least, to the extent a two-dimensional surface can approximate high dimensional data. As illustrated by Figure 11, dense areas of space attract numerous SOM neurons. This property leads each SOM neuron to contain approximately the same number of data points: the SOM evolves into a crude histogram, which allows the user to gain an understanding of the proportion of various patterns in the data.

There are critiques of the SOM. Bishop et al (1997) point out that the SOM is based on heuristic rules, not on rigorous mathematics. They prefer modelling the data with mixture models identified by the Expectation Maximization (EM) (Dempster et al 1977) algorithm. Also Svénsen (1998) proposes a rigorous alternative to the SOM: the Generative Topographic Mapping (GTM). Another type of critique is concerned about the accuracy of topology preservation: Flexer (1999) proposes that the SOM should be replaced with the combined use of k-means clustering and Sammon mapping. This thesis acknowledges those critiques. As stated above, the methodology of this thesis does not require the user to use the SOM; he can select any method that allows understandable study of data, and labelling of the data points serves the tools of variable selection and feature extraction. This thesis uses the SOM as it is widely adapted for visualization of data (Kohonen 2001), and serves exploratory data analysis (Vesanto 2002).

3.2.2 Visualizing the trained SOM

Visualization of the SOM is based on visualising the graph illustrated in Figure 11. The simplest visualization is the U-matrix (Ultsch 1993), presented in Figure 13, which shows how stretched the edges connecting the graph nodes are. The vertical bar in the right of the plot provides the key for reading the distances between the nodes. Such visualization illustrates the cluster structure of the data. For example, Figure 13 suggests that the data contains four clusters, one in each corner of the SOM, and that

the cluster bottom left is best separated from the other clusters. This is confirmed by Figure 11.

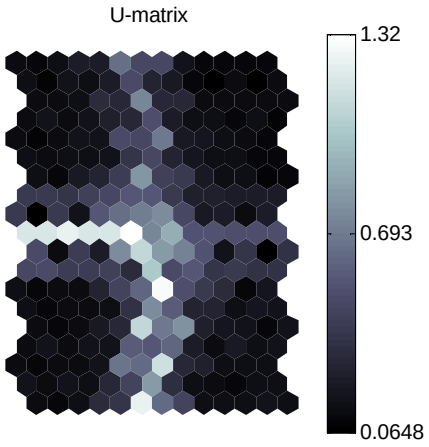


Figure 13. The U-matrix that shows the distance between the nodes of the graph of Figure 11

The next step is to study, where each of the SOM nodes resides in the data space. This is done with component planes. For example, consider the bottom left corner of the SOM described by Figure 13. The position of this corner is presented by the component planes of Figure 14: as the colour of this corner in the left component plane 'dim1' is dark, the value is low, approximately -1.25. Similarly, the value of variable 'dim2' is light, which corresponds to the value of approximately 3.7. This confirms the earlier assumption that this part of the SOM is the most separate cluster in Figure 11. Take another example: the top left corner has high values of the 'dim1' variable, and average values of the 'dim2' values. The number of component planes equals to the dimensionality of the data. Conceptually, the study of any-dimensional data is as easy as the study of the presented two-dimensional case.

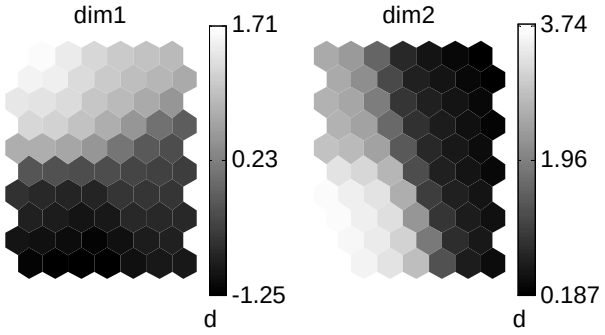


Figure 14. The component planes showing the locations of the neurons of the SOM of the Figure 11

The analysis of various areas of the SOM is supported by clustering the SOM mapping using the method of Vesanto and Alhoniemi (2000). Figure 15 illustrates the clustering of the SOM of Figure 11 to four clusters. Such analysis supports discussion (e.g. Siponen et al 2001), as they can e.g. discuss what are the characteristics of the cluster four, instead of referring it with the vague expression of lower left area of the SOM.

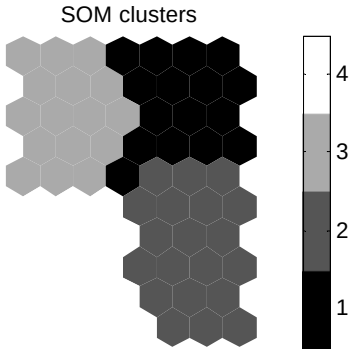


Figure 15. A clustered SOM

From the viewpoint of this thesis, the most important benefit of SOM clustering is labelling of the X-data. Consider a vector of Y-data and introduce it to the SOM. One of the SOM units is closest to this vector: this is the Best Matching Unit (BMU). The BMU already belongs to a cluster as illustrated by Figure 15. The considered vector of Y-data is labelled with this label. And the label can be transferred to the X-data, as discussed in the subsection 2.4.2. Note, that this approach to clustering is not as faithful as clustering the original data: the SOM mapping distorts the original data, and the clustering is based on this distorted mapping. However, the SOM based clustering allows understandable characterisation of the various SOM areas, for example, characterise the cluster one of Figure 15. When analysing high dimensional data, this is more understandable than looking at a matrix of cluster centroids.

As discussed in the previous section, the SOM surface serves as a histogram of typical patterns. Consider, for example, the right subplot of Figure 14. While most of the values of the variable ‘dim2’ are low, close to 0.2, approximately a quarter of the points manifest high values close to 3.7. This support the analysis of a given data sample: whether the value of a variable is high or low, or whether a pattern of data is typical or atypical.

The SOM supports correlation analysis. Consider, for example, the SOM of Figure 16, which illustrates an eight-dimensional space elaborated in chapter 6. Look at the general resemblance of the component planes: the variables form two groups. *Var1*, *Var2*, *Var3* and *Var 7* comprise one group, while variables *Var4*, *Var5*, *Var6* and *Var8* comprise another. The similarity of component planes indicates that the variables are well correlated. Also note that *Var4* is a negative image of the other variables of this group: it is negatively correlated to the three other variables. While such results could be achieved by standard correlation analysis, the SOM offers non-linear analysis. Consider for example the cluster 6 defined by the *SOM clusters* plane at the bottom right part of the figure. This is an area where the general correlations do not hold: *Var3* is not high, but *Var1* and *Var2* are. Also *Var6* is unusually low. This may intrigue the analyst.

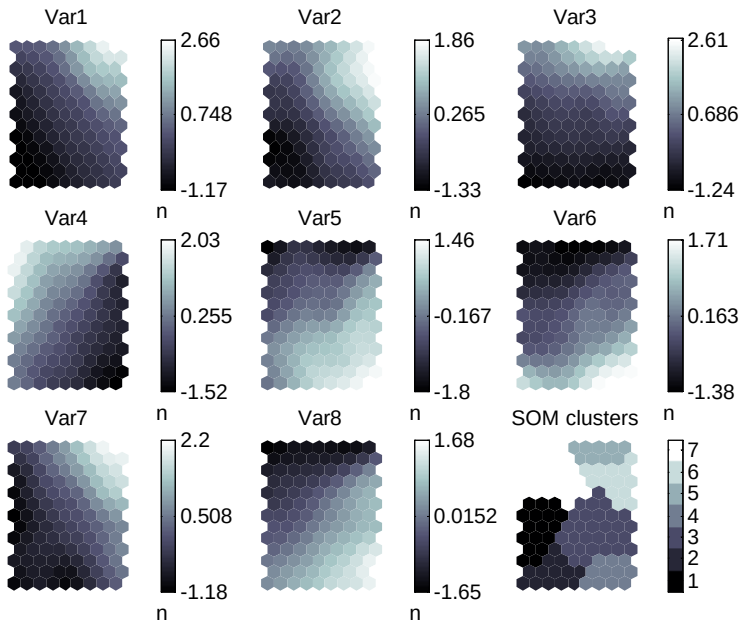


Figure 16. A SOM illustrating an eight dimensional space

Another basic SOM visualization is the hit-histogram, illustrated in Figure 17. It shows how a set of data is quantized onto the SOM. To calculate the hit-histogram, each point is introduced to the SOM, and the BMU of each point is calculated. The counters at each SOM-node are incremented by as they gain BMU nominations. The hit-histogram illustrates the final values of these counters. The Figure 17 shows how the four clusters of Figure 11 land on the SOM illustrated e.g. by Figure 14; the larger the dot, the higher the counter count.

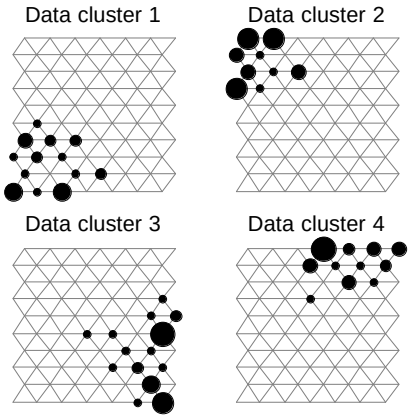


Figure 17. The hit-histograms illustrating the positions of the four clusters of the training data

An application created for on-line use at the Hitura concentrator is presented in Figure 18. The underlying SOM is a mapping of the process feed types. The dark ‘worm’ is the state of the feed type. The path shows the evolution of the feed type from the area of '*poor-talc-amp*’ feed type towards the area of ‘*rich-talc-amp’*.

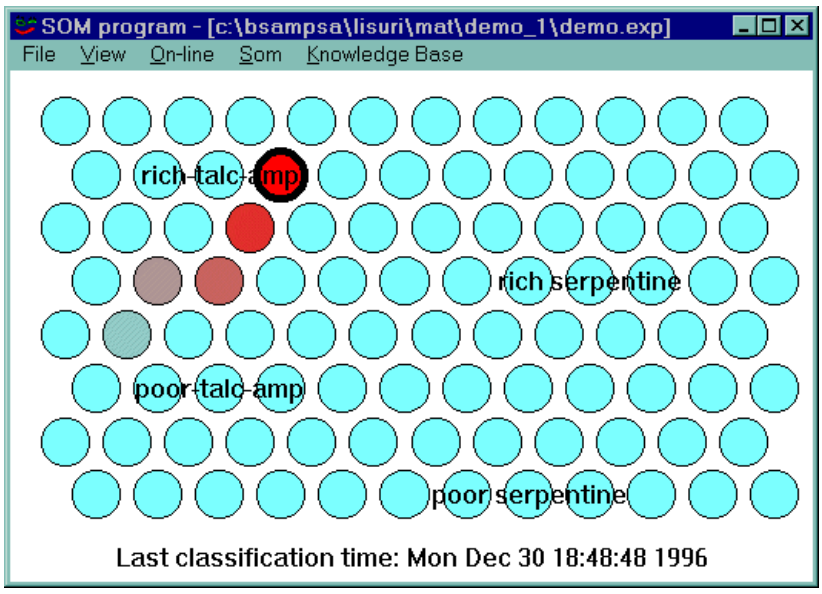


Figure 18. The user interface created for the Hitura concentrator to illustrate the development of the process feed type

Extended discussion of the use of the SOM in visualization is offered, for example, by Himberg (1998), Vesanto (1999, 2002), and Vesanto and Alhoniemi (2000).

3.3 Summary and discussion

Visualization is a thoroughly studied field of science. This chapter discussed basic algorithms using the three criteria of this thesis. Supervised operation is preferred as it allows the visualization to emphasise the aspects defined interesting by the user. This thesis uses the simple scatter plot augmented with icons that include the label information. Unsupervised visualization, based on the SOM, is used in this thesis to support the user's task of defining the problem. This chapter also discussed the robustness of the visualization tools intended for the study of high dimensional data. The scatter-plot, in its simplicity, is robust. Also the SOM is robust when compared with the PCA or FDA, which are disturbed by outliers. Considering the scope of this thesis, understandability is a central issue. The scatter plots are directly understood by process engineers and operators.

The understandability of the SOM among process engineers may be a problem. Still, this thesis suggests the use of the SOM as the clustered SOM well supports labelling of the data to reflect the user's conception of the problem. The process control staff may be trained with the approach of this chapter: to first describe the SOM as a mapping of typical data patterns, e.g. process states; and to possibly show the U-matrix. After establishing the mapping aspect, the teacher can introduce the component planes, and then introduce the other visualization tools as necessary. If trained properly, the process engineers seem to accept the SOM. For example, the Hitura staff accepted the on-line tool of Figure 18.

4 Variable selection

Variable selection, sometimes also called feature selection, is an important aspect of data analysis. The importance of variable selection is manifested by the learning models of chapter 2: all of them describe the selection of data as a major step of analysis. Zooming analysis onto relevant information supports both creation of mental and mathematical models; irrelevant information confuses the learner.

Due to its importance, variable selection is a well-established field of statistical study; e.g. Dash and Liu (1997) have studied how to select the variables that best serve classifiers. Still, algorithmic variable selection, according to my experience, does not prevail in industry. This thesis assumes that engineers do not find algorithmic variable selection useful. Variables are selected manually, which is laborious: the automation system offers hundreds of signals. This thesis looks for a variable selection approach that seeks the approval of the engineers. The approach is supervised: it returns variables relevant to the user's problem; the approach is robust: the only required step of pre-treatment is scaling, e.g. to unit variance; and the algorithm is understandable, in principle, in parameterisation, and in presentation of the output.

The scope of the approach of this thesis is Statistical Pattern Recognition (see e.g. Jain et al 2000), as the goal is to create accurate classifiers. This chapter starts by a general introduction to algorithmic variable selection, then surveys the field of Statistical Pattern Recognition from the viewpoint of variable selection, and proceeds to describe the approach of this thesis.

4.1 Algorithmic variable selection

Algorithmic variable selection is usually based on combinatorial optimisation. The algorithm looks for the set of variables that minimises the cost function, see Langley (1994) for a survey. Langley illustrates the task of variable selection as in Figure 19, which shows all combinations of four variables. At left, the set is empty; at right, it contains all the four variables. The goal is to browse this graph to find the best box.

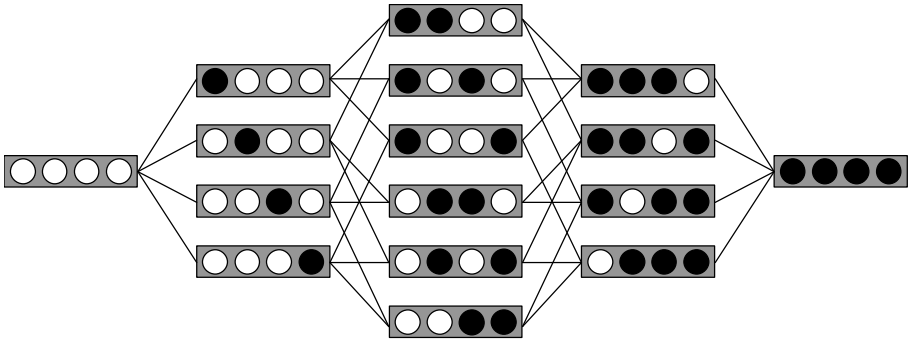


Figure 19. The graph of all possible four variable combinations

Langley defines four dimensions along which variable selection techniques differ. The first is to decide from which box of Figure 19 to start from. Two common approaches (see e.g. Aha and Bankert 1996) are feed forward selection and backward elimination approaches; in former, the user starts with an empty set and starts adding variables; in latter, the user starts with all variables, and starts eliminating them.

The second aspect of Langley is to solve, which path to take when studying various boxes of Figure 19. The simplest approach is to exhaustively study all of them; however, this requires 2^n evaluations of the cost function, which is unfeasible with high-dimensional data sets. If the study of one box takes a second, and the size of the studied variable set is thirty, the search takes 34 years. A simple resolution is provided with the feed forward selection, in which the search starts with an empty set, and variables are added to the set until the predefined end criterion is met. This thesis prefers the more advanced algorithms of combinatorial optimisation, and uses the Tabu-search (e.g. Kreher 1999) to browse the space. It resembles feed forward selection, but has some capability of escaping the local minima (see Appendix A). Ferri et al (1994) compare greedy algorithms with the Genetic Algorithm (GA), and suggest that their combination may best support variable selection. Also Vafaie and de Jong (1993) advocate the GA. This thesis does not use GA, as the used Tabu-search algorithm provides adequate results, and is faster than the more randomized GA approach. The ability to provide answers in minutes, rather than in hours, is an important aspect of usability from the viewpoint of an engineer.

The third viewpoint of Langley is the cost function that is the basis of the decisions of the optimisation algorithm. In statistics, the common cost function is the accuracy of the created statistical model. The cost function can, for example, aim for the most accurate regression model (George and McCullough 1993, Miller 2002); or aim for the most accurate classifier (Koller and Sahami 1996, Dash and Liu 1997). This approach, discussed e.g. by Jain et al (2000), is the basis of the methodology of this thesis. In addition to finding the most accurate model, it can try to regulate the number of variables in the set, for example by penalizing the accuracy by the number of variables in the set. The cost function is the core of optimisation; this discussion is elaborated in the next section.

The fourth viewpoint of Langley is the stopping criterion of optimisation. Termination of the optimisation is partly defined by the cost function, and partly by the optimisation algorithm that studies the developments of this cost function. The conflicting goals are to provide an unbiased answer, e.g. to avoid local minima, and to ensure efficient performance of the variable selection tool.

4.2 Survey of classifier accuracy based cost functions

This section surveys algorithms that look for variables that create the best classifier. Classifiers can be created based on a number of theories, created e.g. in the field of Statistical Pattern Recognition (SPR). The SPR is the basis of this section, and more specifically, the SPR-taxonomy of Jain et al. (2000) presented in Figure 20.

The most accurate classifiers are based on the Bayes decision theory, depicted in the leftmost box of Figure 20. They are supervised and robust, and allow understandable

representation of the results. In this sense they are the optimal tools for variable selection. The Bayesian approach is discussed e.g. by Charniak (1991), Buntine (1996), Haddawy (1999), and Jordan et al (1997). A simplified Bayesian approach, called the Naïve Bayes, is discussed e.g. by John and Langley (1994), and Domingos and Pazzani (1996). However, the handicap of the Bayesian approach, as discussed by Gaag et al (2002), is the laborious definition of the class-conditional densities of the studied variables. Industrial engineers have neither the statistical expertise nor the resources for creation of the required Bayesian models of their process.

This section looks for the alternatives of this optimal approach. The study starts by the algorithms at the left of Figure 20, and proceeds to the right. The algorithms at the left require rigorous models of the studied domain, but return sophisticated results of analysis; when advancing to the right, the workload of the engineer gradually reduces to the level of only providing the data. At the same time, the algorithm is less and less equipped to return relevant results. This section looks for the compromise that best serves the methodology of this thesis.

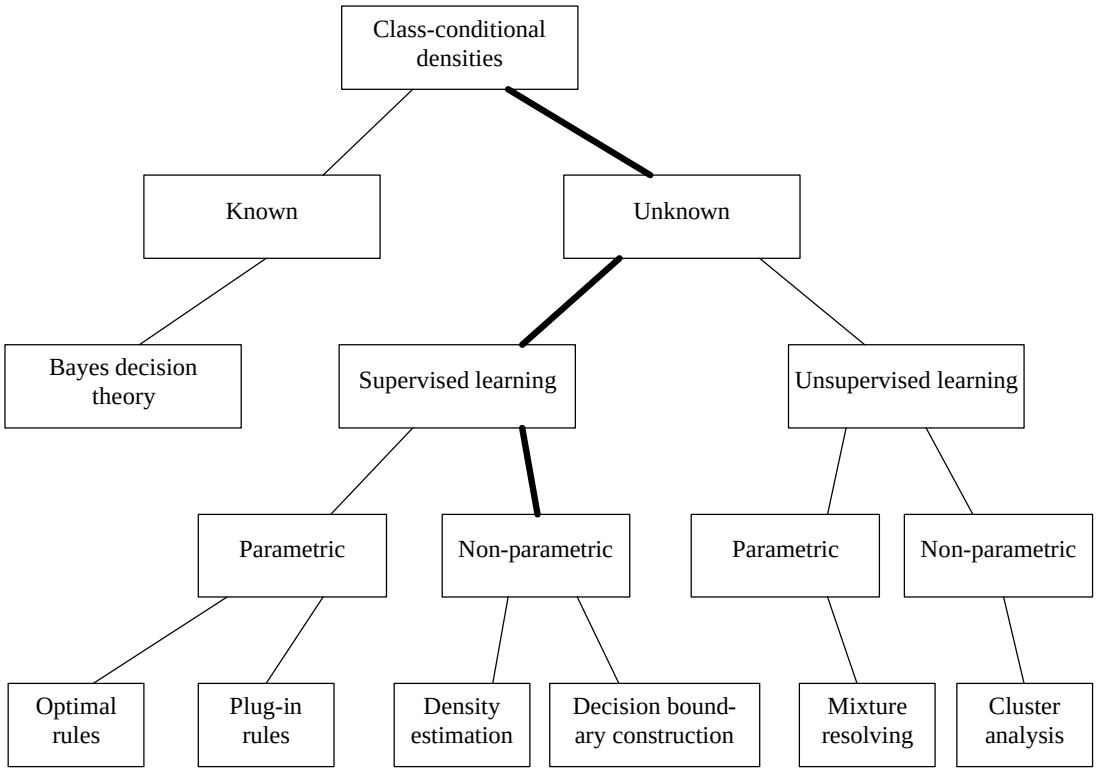


Figure 20. Various approaches in statistical pattern recognition (Jain et al 2000). The approach of this thesis, non-parametric supervised learning, is illustrated with the thick line.

The remainder of this section studies the right hand side sub-tree of Figure 20, which assumes the class-conditional densities to be unknown. The models learn these densities from data. In this branch, the first dichotomy is to supervised and unsupervised learning. As discussed in the previous chapters, supervised algorithms have the benefit of returning information defined interesting by the user, but have the drawback of requiring the user to explicate his interests. Still, the supervised learning is promoted in this thesis, as the industrial engineer is faced with the hundreds of variables stored by the automation system. Unsupervised tools may return statistically significant results that are not related to the studied problem. Such results may confuse the user. Advocating supervised operation is the main tenet of this work.

The next dichotomy, in both supervised and unsupervised branch, is to parametric and non-parametric algorithms. Parametric methods require the user to know, or assume, some properties of the probability distribution; for example, model the clusters with variance-covariance matrix, which assumes the clusters to be Gaussian. The benefit of parametric models is compactness: the number of parameters is usually low compared to parametric models. Also, the parametric models allow rigorous mathematical analysis. Their problem is unrobustness that follows if the data violates the assumptions; for example, if non-Gaussian clusters are modelled with Gaussian tools. This thesis advocates non-parametric models, which make less assumptions of the shape of the data. This makes them more robust. The remainder of this thesis discusses the various degrees of supervised and parametric nature by looking at the algorithms listed as leafs in the bottom of Figure 20.

The first leaf implements supervised parametric learning based on Bayesian *Optimal rules*, created e.g. by the rigorous work of an expert. This work results in a thorough model of the domain, which allows high accuracy classifiers, and comprehensive analysis of the sources of uncertainties. However, in my opinion, such rigorous modelling is feasible only in research projects: it requires extended and explicated knowledge of the studied process, understanding of statistical mathematics, and, usually, a few months of labour. Such resources are seldom found in industrial plants.

A simpler variant of the Bayesian approach is the creation of *Plug-in –rules*. This method uses simpler distribution types, and aims to automatically identify them from data. A typical simplification, according to Jain et al, is to assume Gaussian distribution in the clusters. A further simplification is to assume each cluster to have identical shape. This approach maintains a high degree of supervision, and adds feasibility, as the models can be learned from data. However, the strong assumptions of the shape of the data lead to biased results if the data does not comply with the assumptions of the algorithm. Even these biased results can be useful to the user, if he can assess their reliability, and understand how they may be biased. For example, consider a set of data that contains relevant information, and one outlier. The Fisher Discriminant Analysis may be biased by this outlier, and claim that there is no information in the data; if the user finds the outlier, understands it influence to FDA, and removes it; he may gain useful results. However, this requires statistical skills.

Non-parametric algorithms are better suited for the analysis of raw industrial data: they make fewer assumptions of probability distributions. For example, they are less vulnerable to outliers. The first of the two leafs is *density estimation*, realized e.g. by mixture models (e.g. John and Langley 1995) or data binning (Domingos and Pazzani 1996, Liu and Setiono 1997). Bonnlander and Weigend (1994) compare the properties of these two approaches. One derivative of binning is the Rough Set theory, proposed by Swiniarski and Skowron (2003). These algorithms are similar to the above algorithms of parametric modelling, as they aim to resolve the probability distribution of the data.

Decision boundary construction is also a supervised, non-parametric approach, but instead of modelling the probability distribution, these methods look for boundaries that can separate the clusters. A simple such algorithm is the decision tree, described e.g. by Murthy (1997). Another simple approach is the Learning Vector Quantization (LVQ) created by Kohonen, and illustrated e.g. by Ray and Chan (2001). A contemporary approach to decision boundary construction is the Support Vector

Machine (SVM) created by Vapnik. See Burges (1998) for a tutorial or Weston et al (2000) for using SVM in variable selection.

The above approaches are still more complex than required by the variable selection tool of this thesis. The algorithm of this thesis is based on a very basic classifier, the K Nearest Neighbour (k-NN) algorithm (e.g. Dasarathy 1991). The k-NN is widely used in practice (Laaksonen and Oja 1996). While its drawback is computational cost (Skalak 1994), its robustness and understandability are appealing. Vapnik (1998) advocates simplicity by stating: "If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem". The k-NN is a simple approach that is adequate for this thesis.

Two unsupervised techniques are studied for the completeness of discussion, and because unsupervised algorithms (e.g. clustering) are used in the first phase of the methodology of this thesis, the phase of problem definition. First of these algorithms is mixture resolving, used to estimate probability densities (e.g. Parzen windows 1962). These estimations can be used in unsupervised variable selection. Dash et al (1997) show how to select variables based on entropy measure; Yand and Moody (1999) and Back and Trappenberg (2001) show how to select variables based using independent component analysis (ICA) (e.g. Hyvärinen et al 2001). Yang and Pedersen (1997) compare unsupervised variable selection methods in the domain of text categorization. However, these unsupervised methods do not know the goals of the user, but return the statistically most significant results. If they are not related to the user's problem, they confuse the user.

The second of the two unsupervised algorithms is cluster analysis. Clustering looks for natural groupings of data. Montanari and Lizzani (2000) show how to select the variables that best contribute to the clustering of the data. Lagus et al (2001) show how cluster variables according to their similarity. This thesis applies clustering onto the results of SOM based vector quantization, as discussed in subsection 3.2.2.

The above study of Statistical Pattern Recognition is augmented by two concepts: size of the selected variable set, and the dichotomy between filters and wrappers.

In variable selection, the user may want to favour the small size of the returned variable set; the benefits are discussed e.g. by Back and Trappenberg (2001). Small sets can be achieved e.g. by penalising large variable sets, as is done with the Akaike's Information Criteria (AIC) (e.g. Akaike et al 1998) or the BIC criterion (Schwarz 1978). More recent methods are bootstrap (Efron 1993) and cross-validation (Devroye 1996); see Kohavi (1995) for a comparative study. According to Madigan and Rafferty (1994) the theoretically most justifiable is the Bayesian algorithm. This thesis is based on the simple approach of restricting the number of variables output by the variable selection tool, most often to two, as two-dimensional spaces are straightforward to illustrate with the scatter-plot.

Another issue relevant to variable selection is the dichotomy to wrappers and filters; see Kohavi and John (1997). The wrapper approach explicitly considers the needs of the final application when selecting the variables. For example, Publication P7 shows how to find variables that train the SOM, which best illustrates the phenomenon of interest. The filter approach is simpler: variables are searched based on general statistical measures: for example Publications P6 and P8 look for cluster separability,

and do not consider the application that might be build on the found variables. According to Kohavi and John, the benefit of the wrapper approach is to surface variables that serve the final application; and their drawback is high computational load, as the final application must be included in the cost function, which may be evaluated thousands of times when looking for the best set. The methodology of this thesis is based on the filter approach.

4.3 The approach of this thesis: k-NN based study of cluster separability

The cost function of this thesis looks for the variable set, which spans a space, where the clusters defined by data point labels best separate. These labels have been defined by the user to explain what he considers interesting. The goal of variable selection is to find variables that share information with this phenomenon of interest. Two variable sets are illustrated in Figure 21. The bivariate set on the right contains interesting information: the clusters defined by the user's labels separate. These variables can be further studied to learn how the phenomenon of interest manifests in the studied data, or the variable pair could be the basis of a classifier application used to label new points of data.

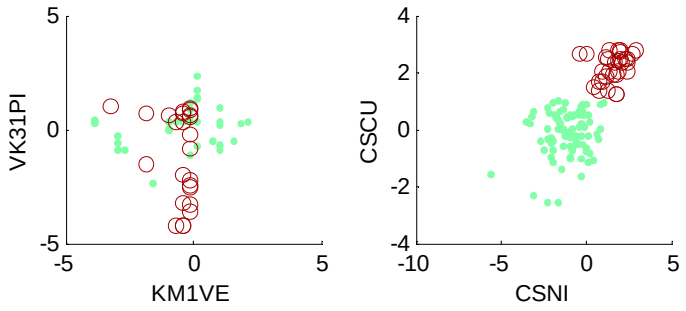


Figure 21. A high-dimensional data illustrated by two variables pairs; the pair at the left is randomly selected; the pair at the right is the pair that best separates the clusters

This thesis uses combinatorial optimisation to browse the variables. Optimisation requires a cost function. The publications of this thesis have examined three variants: the publication P6 studied parametric measures, the publication P7 studied wrapper methods, the publication P8, and this thesis, uses the non-parametric method elaborated below.

Figure 22 illustrates the cost function of this thesis. The cost function traverses all data points, and collects the K nearest neighbours of each point, as depicted by the circle in Figure 22. The function then calculates the proportion of data points that share the label of the studied data point, surrounded with the box at the centre of the circle. If these neighbourhoods, on average, are uniform, the clusters separate. For example, most of the data points of Figure 22 have such uniform neighbourhoods.

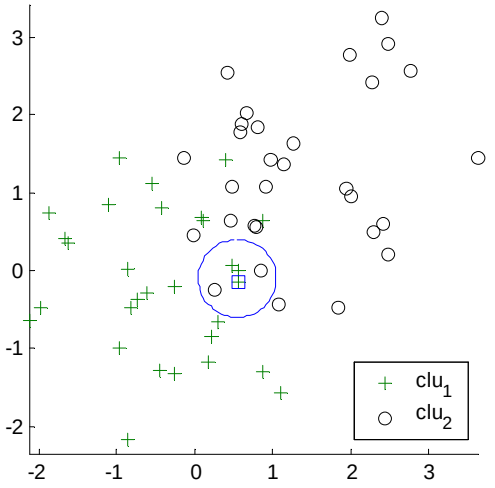


Figure 22. A data point with a neighbourhood of four points

The cost function is formulated mathematically in Equations 4 and 5. The nominator of Equation 4 counts the points x_j that belong to the neighbourhood N_i of the studied point x_i, and share its label c_n. The neighbourhood comprises the K nearest neighbours, measured by the Euclidean distance. This count is divided by the size of the neighbourhood, K. When all points have been studied, the Equation 5 calculates their average, which produces the separability measure, C , that is the cost function used in this thesis. The separability takes values from zero to one; values close to one imply uniform neighbourhoods, that is, good cluster separation.

$$p_i(c_n \mid x_i) = \frac{\#\{x_j \mid x_j \in c_n, x_j \in N_i\}}{K} \tag{4}$$

where,

x_i	=	the studied point
c_n	=	the label of the studied point
x_j	=	other points of the data set
N_i	=	the neighbourhood of point x_i
K	=	the size of the neighbourhood

$$C = \frac{\sum_{i=1}^{m} p_i}{m} \tag{5}$$

where,

p_i	=	the uniformity around data point x_i
m	=	the number of data points
C	=	the separability of clusters

The cost function has one parameter, the size of the neighbourhood K. Values of K close to the number of data points do not perform well: if all data points are included

in the neighbourhood, all variable combinations return the same cluster separability. If the value is too small, the separability measure advocates local separability of clusters, and ignores global separability of clusters. This may lead to a chess-board type separation of clusters. The heuristic rule applied in the cases of this thesis is to take the median cluster size and divide it by two.

The above method is sensitive to scaling of the variables. The Euclidean distances are dominated by variables with high interpoint distances; roughly speaking, the variables with comparably high variance dominate the analysis. The user can use this to intentionally favour some variables, see e.g. Hätönen et al (2003). The case study of this thesis uses the straightforward approach of scaling all variables to unit variance. Another statistical notion is that the above cost function does not measure distances between the label types: the labels of the neighbours are identical to the one at the centre of the neighbourhood, or they are different. A possible spin off of this thesis would be to include distances between labels into the above cost function.

The most attractive feature of the above cost function is that it makes no assumptions of the shape of the clusters. This is illustrated in Figure 23 that illustrates a synthetic data comprising three label types (Clu_1 – Clu_3). Most of the data points are bundled at the left, but at the right, there is some separability between the data points carrying different labels. The non-parametric cost function of this thesis captures such partial separability. Traditional statistical tools look for global separability of clusters, and probably would ignore the local separability of Figure 23. Similar benefits can be achieved with tools based on non-parametric mixture modelling, discussed in the previous section. However, the tool of this paper offers similar results with a simpler concept.

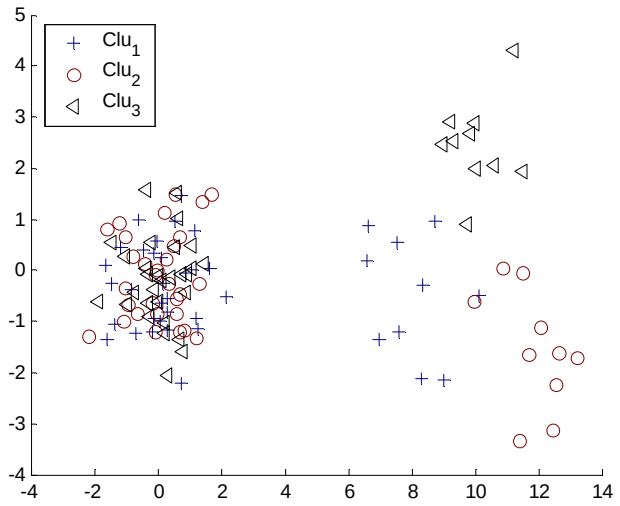


Figure 23. Three clusters (Clu_1-Clu_3) that are generally mixed, but do separate at the right hand side of the plot

A major drawback of the cluster separability measure of this thesis is computational complexity. The algorithm requires the calculation of all interpoint distances, which amounts to *n*(n-1)/2* calculations, where *n* is the number of data points. According to my experiences, the cost function of this thesis becomes unfeasible, if the number of data points exceeds approximately one thousand. This is not an unsolvable problem: the user can reduce the number of data points by e.g. random sampling or with vector

quantization (see e.g. Györfi 2002). This problem applies to the number of data points; the number of variables is less of a problem: the computational load increases only linearly with the number of variables.

The above cost function does not make assumptions of the shape of the clusters, but like all data driven statistics, it assumes the data to adequately represent the problem. Problems can be caused e.g. by a data set where the number of data points per cluster is small, say, less than ten. In such case random effects may bring them close to each other and the algorithm may return a biased variable set. Classic statistics solves this problem by defining the statistical significance of the results, for example, with p-values. While the non-parametric approach of this paper does not allow the calculation of p-values, it allows the use of other methods, such as cross validation (Devroye 1996).

The benefit of the above cost function is its compliance with the criteria of this thesis. It allows supervised operation, as the cluster separability is measured based on the labels defined by the user; it is robust, as it makes no assumptions of the shape of the clusters; and the approach is understandable because its concept and parameterization are simple.

4.4 Summary and discussion

Variable selection is an important phase of data analysis: it is included in all learning models described in chapter 2. Variable selection is especially important in the study of industrial data, where most of the hundreds of variables offered by the automation system are irrelevant to the user's problem.

This chapter surveyed the field of Statistical Pattern Recognition to illustrate various approaches to variable selection. While the Bayesian approach provides the most accurate algorithms, it requires elaborated modelling of the studied domain. Simplified approaches identify the required probability distributions from data, but, if based on parametric models, make strong assumptions of the data; and if based on non-parametric models, require elaborated computation to estimate the probability density functions. All approaches based on explicit identification of the probability density functions are influenced by the curse of dimensionality: the problem of reliably creating high dimensional models with limited data.

The approach of this thesis is not based on estimation of the probability density function. It uses the simple k-NN based approach to study the neighbours of each data point. If these neighbourhoods contain data points with uniform labels, the clusters separate. Such a variable set contains information that is relevant to the problem defined by the user's labels. While the method of this thesis may not be as accurate as e.g. elaborated Bayesian classifiers, the approach of this thesis excels in robustness and understandability. The approach of this thesis does not require explicit identification of the probability density functions.

5 Feature extraction

A definition of the Merriam-Webster's Collegiate Dictionary for the word *feature* is ‘a prominent part or characteristic’. According to the constructivistic theory, humans extract features automatically and constantly. For example, a process expert knows to look, instead of the absolute values of variable A, to the difference in value of variables A and B. Or he may know to look at the trend of the signal C instead of its absolute value. While variable selection works with raw values of data, feature extraction manipulates the data to surface its relevant aspects.

The statistical community acknowledges the feature extraction as part of data analysis: in the CRISP model (Figure 3, page 6), feature extraction is part of the data preparation phase; in the KDD-process (Figure 4, page 7) feature extraction is part of the data transformation phase. Feature extraction may serve the above goals of surfacing relevant information, or it may serve computational aspects by reducing the dimensionality of data to manageable level (see e.g. May and Jain 1995).

The angle of interest of this thesis is the surfacing of relevant information. As with variable selection, the state-of-the-art of industrial feature extraction seems to be manual selection of the feature extraction algorithm and creation of the model. Even the statistical literature does not offer general approaches for finding the best feature extraction algorithm for the task at hand. Exceptions are the approaches of Jaimes and Chang (2000) and Lindner and Studer (1999): they algorithmically select the feature extraction algorithm that creates good models for the given case. This is also the approach of this thesis. The goal is to offer the engineers a library of potentially useful algorithms, and a tool that selects the algorithm that serves the task at hand. The user does not have to know beforehand how e.g. the Fourier transform operates; the user can get acquainted with this algorithm, if it proves to extract interesting information from his data.

This approach supports learning. The tools of this chapter create simple models that illustrate the problem more clearly than a set of original variables. The user learns by looking at the models that extract this useful information. If the user accepts these mathematical models, he internalises them as mental constructs used to read data. Knowing what to look for in a set of data is a hallmark of a process control expert.

5.1 Survey of feature extraction algorithms

A basic algorithm of feature extraction, according to the survey of Jain et al (2000), is the Principal Component Analysis (PCA). It defines a feature space of principal components that aims to capture the variance of the data (see e.g. Figure 7, page 16). The projection pursuit algorithm by Friedman and Stuetzle (1981) performs a similar task; the difference is that projection pursuit aims to capture the non-Gaussian information of the data. Another algorithm for dimension reduction is the Sammon mapping (Sammon 1969), which provides a non-linear and non-parametric service. Lerner et al (1999) provides a comparative study of these techniques. A more complex

approach, introduced by Fisher and Principe (1998), uses general information theoretic definitions.

A common denominator of the above algorithms is their unsupervised nature. A linear supervised algorithm is the Fisher Discriminant Analysis (FDA) (Fisher1936) illustrated in Figure 7, page 16. A non-linear approach based on Artificial Neural Networks (ANN) is proposed by Lowe and Tipping (1996). Torkkola (2001, 2002) propose an approach that has information theoretic justification. These supervised approaches can extract information that illustrates the problem defined interesting by the user.

Unrobustness is a problem in the study of industrial data, also in the extraction of features. Algorithms based on strong assumptions are unrobust if the data fits these assumptions; for example the PCA and the FDA are based on the variance-covariance matrix, and are unrobust towards outliers. These two algorithms also illustrate another typical limitation of feature extraction algorithms: linearity. The linear models may fail to capture non-linear phenomena. Strong assumptions often allow mathematical tractability and computational efficiency. However, they require the user to acknowledge these limitations: to pre-treat the data to serve the method, or to understand the sources of bias in the results. This thesis advocates non-parametric methods that make fever assumptions of the data.

Understandability is as important in feature extraction as it is in visualization and variable selection. Learning from features is possible only if the feature extraction models are understandable. A counter example are the black-box models created e.g. by Artificial Neural Networks. These models are accurate but do not allow human analysis and learning of the models, see Publication P8 for elaboration. An interesting approach in the sense of understandability is Gene Expression Programming (e.g. Ferreira 2001), which uses Genetic Algorithms to formulate mathematical equations that extract useful features. These features capture complex phenomena, but are understandable only to mathematically literate persons. The method of this thesis also extract mathematical models, albeit simple ones.

5.2 The approach of this thesis: automatic creation of simple features

The feature extraction approach of this thesis is based on the variable selection algorithm presented in the previous chapter, but requires some additional steps to be taken: the user needs to select a set of feature extraction algorithms and variables to be tested. In practice, the algorithms are selected from a library, and variables are selected manually by the user, or with the variable selection tool of the previous chapter. The process of feature extraction is illustrated in Figure 24. The computer runs all candidate variables through all candidate algorithms, which produces a large number of candidate features. Mathematically, these candidates are vectors, just like the original variables; this allows the use of a variable selection algorithm to look for the feature set that best describes the problem of interest.

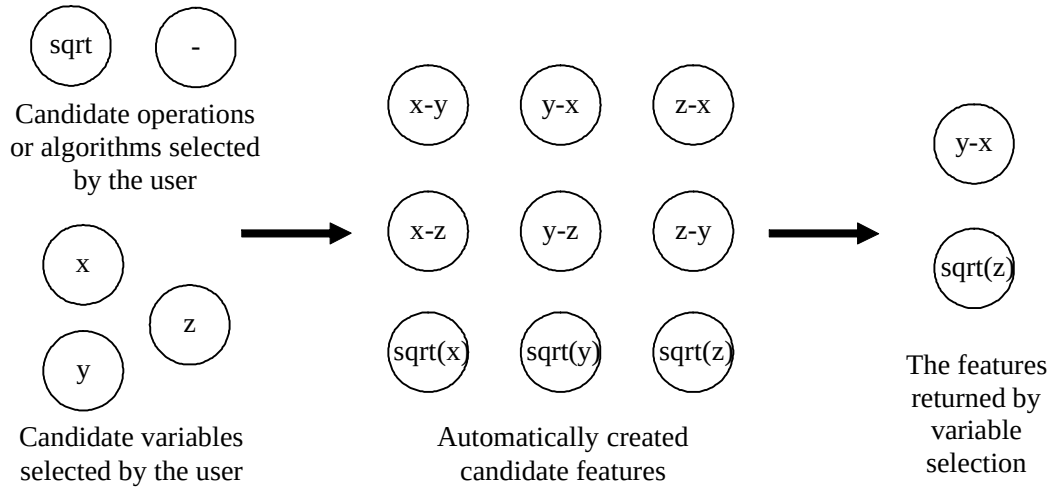


Figure 24. Finding the feature pair that contains information of the interesting phenomenon

This approach complies with the three viewpoints of this thesis. The process is supervised, as it is based on the user's labels. The approach is robust towards outliers and other anomalies in the cluster shapes, as it is based on the variable selection tool of the previous chapter. And the approach is understandable, as its principle is simple. The approach of this thesis is compared with standard statistical algorithms of feature extraction in Publication P8, and illustrated by the Hitura case in Chapter 6.

5.3 Summary and discussion

Feature extraction is an important step of mental and algorithmic study of data. Features are extracted to surface information for further analysis. This chapter showed how to use variable selection to extract features, which robustly and understandably illustrate the phenomenon of interest. The user, by studying the feature extraction models, learns how to study the data to expose information of interest. As automatically selected variables, these features may surprise the learner, and lead him to create new hypotheses that lead to a new iteration of the wheel of learning.

6 Experiments: an industrial case

This chapter illustrates the methodology of this thesis by analysing the case of the Hitura mine. This chapter is organised according to the CRISP data mining model (see Figure 3, page 6). Some data are presented in scaled form to protect the proprietary rights of the Outokumpu Finnmines.

6.1 Business understanding

The Hitura Mine is a nickel mine located in central Finland. The mine capacity is 530.000 t/a ore. The Hitura ore contains 0.65 weight percent (w-%) of Nickel (Ni) and 0.2 w-% of Copper (Cu). It occurs in amphibole-bearing rocks in the contact of a pipe-shaped serpentinite against mica-gneiss. The main sulphide minerals are pyrrhotite, pentlandite and chalcopyrite. The product of the concentrator was in 1995 38 000 t/a of Nickel-Copper concentrates with a grade of 6.5 w-% Ni, 1.7 w-% Cu and 10-12 w-% Magnesium Oxide (MgO). For elaborated studies of the mineralogy, see Pulkkinen et al (1993a) or Ruonala (1996).

The main process control problem at Hitura is the volatility of the feed type (Laine 1997). For example, the proportion of nickel varies between 5 w-% and 13 w-%; the main silicate mineral, serpentinite, varies between 15 w-% and 90 w-%; and the main source of concentrator quality problems, talc, varies between 5 w-% and 13 w-%. In addition to amplitude, also the frequency of variation is high. Major feed type changes occur approximately every other day, minor changes occur many times a day. After each change, the operators must adjust the process for the new feed type. The magnitude of the change is illustrated in Figure 25 with the day-averages of two process control parameters: when the feed type changes from serpentinite to talc-amphibole, these two parameters should be dropped to a fraction of their original values. The process operators do not know the nature of the new feed type: they must adapt the process by trial and error, as depicted by the angled arrow of Figure 25. This takes time, and the found solution may not be in the optimal area.

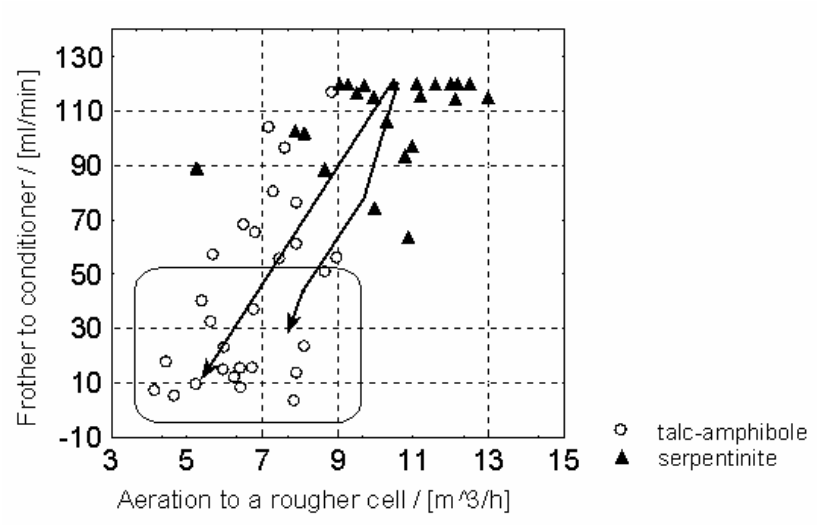


Figure 25. Adaptation from serpentinite to talc-amphibole process feed type

A solution to this problem is presented by the expert system of Publication P2. It provides on-line information of the feed type, and advices how to process it. This system requires two tools: an on-line observer of the process feed type, and a knowledge base of the best practices for each feed type.

6.2 Data understanding

The Hitura mine provides two types of numeric information: off-line data measured by the plant laboratory, and on-line data measured by process instrumentation system. Off-line laboratory data are produced by manual chemical analysis of the samples collected from the main streams of the concentrator. These analyses provide versatile information of the feed type and product quality of the concentrator. The handicap of this information is that it is produced only once per day. The expert system supporting on-line process control must be based on on-line data. This case shows, how to use the off-line laboratory data to define the problem; and how to search for the on-line data for the answers. This discussion is elaborated e.g. in Publication P4.

The off-line variables presented in Table 3 adequately describe the feed type problem of the Hitura mine (Laine 1997). This thesis uses data collected in the period from 19th of March 1994 to 30th of November 1995. After removing the process down times, 526 day average values remained for analysis. This data was used to train a SOM (Figure 26), which became a mapping of the feed types. Training was done with the SOM-toolbox created by Vesanto, Himberg, Alhoniemi and Parhakangas (1999), and with the default training values of this toolbox. The trained SOM was clustered according to the method of Vesanto and Alhoniemi (2000). The number of clusters (seven) was selected heuristically after experimenting with data. The clustering algorithm was hierarchical linkage with the Ward distance measure (e.g. Jain et al 1999). For elaborated discussion of clustering, see e.g. Jain et al (1999) or Li and Biswas (2002). The clustered SOM was used to discuss the feed types of the Hitura mine with the experts of Hitura mineralogy and process control, and was used to label the points of

on-line data with feed type labels. This concluded the problem definition phase of the analysis methodology described in the subsection 2.4.2.

Table 3. The off-line measured variables describing the concentrator feed

Abbreviation	*Variable type*
ONI	Ni content of the ore (w-%)
OCU	Cu content of the ore (w-%)
OS	S content of the ore (w-%)
OFE	Fe content of the ore (w-%)
OFEM	Magnetite content of the ore (w-%)
ONISULF	Ni content of the sulphide phase of the ore (w-%)
OINSOL	Weight-% of residue after assay dissolution of the ore by nitric acid
OMGO	MgO content of the ore (w-%)

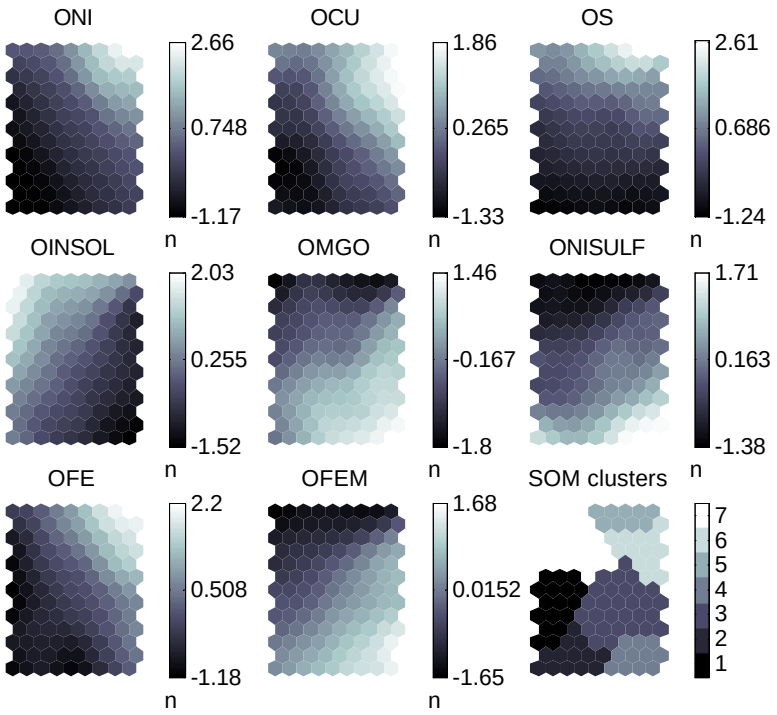


Figure 26. The feed type SOM created with off-line data

The set of data, among which the result must be found, is on-line data: the goal was to create an on-line decision support system for the process operators. The first type of on-line data, available at Hitura, is produced by basic instrumentation: it measures and records air, water and chemical flow rates, pressures, levels in the flotation cells, and set points of the process controllers. Another type of data is produced by the two advanced instrumentation devices: the PSI200 that measures the particle size distribution in process flows; and the Courier 30 X-ray analyser that measures element contents in the slurry flows, for example, the contents of nickel, copper and iron.

To summarise this section: the off-line data is used to define the problem (volatility of the process feed type); and the on-line data is studied to learn how to realise the expert system for operator on-line decision support.

6.3 Data preparation and modelling

The first step of data preparation was scaling the off-line and on-line data to zero mean and unit variance. The process down times were also removed from the data. Other pre-processing, such as elimination of outliers, was neither necessary, nor performed.

The actual data modelling started by creating the mapping of the feed types using the SOM of Figure 26. This SOM provided each day average data point with a feed type label, as illustrated in Table 2 in page 12. To elaborate, consider day Z in history. The off-line laboratory provides the chemical analysis of the day average assay of the process feed; the analysis result allows this day to be positioned onto the feed type SOM of Figure 26. This feed type label is attached to the off-line data point. The automation system data base is queried for the day average result of the same day, and the off-line defined feed type label is attached to this data point. The rest of this chapter is based on analysing such feed type labelled on-line data.

6.3.1 Creation of the on-line observer

The first analysis looked at the 24 signals produced by the Courier 30 (C30) analyser. The objective was to learn, whether the C30 information can be the basis of a feed type observer.

The variable selection algorithm of this thesis was applied to return the variable pair that contains best information of the feed types. This pair is illustrated in the subplots of Figure 27; two plots are used to avoid over-plotting. The plot on the left illustrates clusters 1, 3, 5 and 7; and the plot on the right the clusters 2, 4 and 6. As presented in the figures, the clusters occupy distinct areas in the space: these C30 signals could serve as a basis for creating an on-line observer of the feed type.

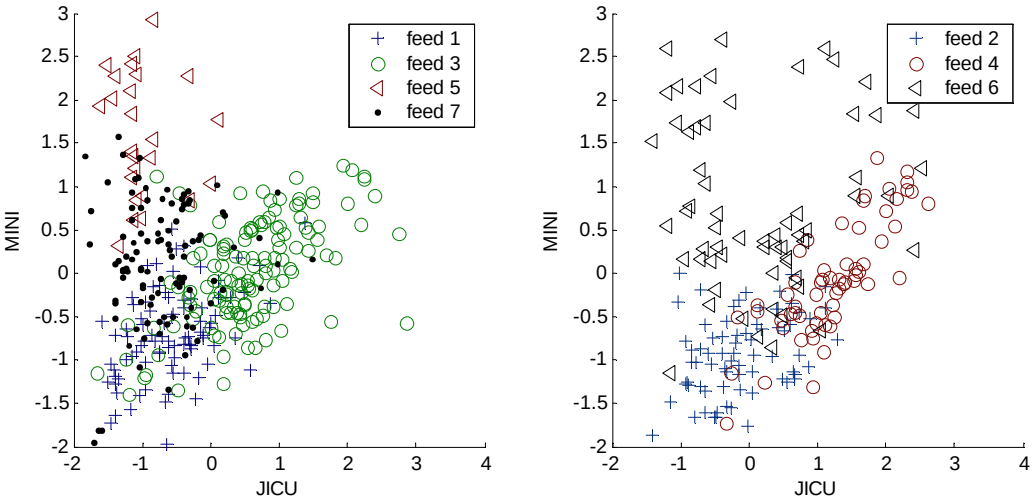


Figure 27. The best two-variable set of describing feed types

The next step is to justify and learn from the findings. The first of the two on-line variables, MINI, is the C30 nickel channel intensity for the concentrator feed. It describes an important property of the sulphide phase of the feed: the nickel content. The silicate phase of the feed is described by the JICU variable, which is the C30 Copper channel intensity for the concentrator tailings (=concentrator waste). The amount of Copper in the tailings indicates the mineralogical conversion of the feed type: the converted form of Copper-mineral, vallerite, has poor floatability and goes to the tailings. In conclusion, the two found variables describe the feed type, and allow mineralogical justification.

However, from a process control point of view, the JICU is not favoured: it is measured from the output of the process, which allows it to be inadvertently manipulated by the process operators. The JICU might not describe the process feed type, but the ability or inability of the operators to react to a change in the feed type. Another reason is timing: when the JICU reacts to a feed type change, all of the process has already been disturbed by the feed type change. The next analysis looks for an alternative for the JICU-variable.

This second study looks for features that are based on data collected from the early stages of the process. The studied variable set comprises twelve variables describing the feed and the rougher concentrate, measured early in the process. Candidate feature extraction algorithms were the sum, difference, product and ratio. Running all variable combinations through all algorithms, and addition of the original variables to the set, produced 276 candidate features. The variable selection tool returned the two-variable set presented in Figure 28. Again, the clusters separate well.

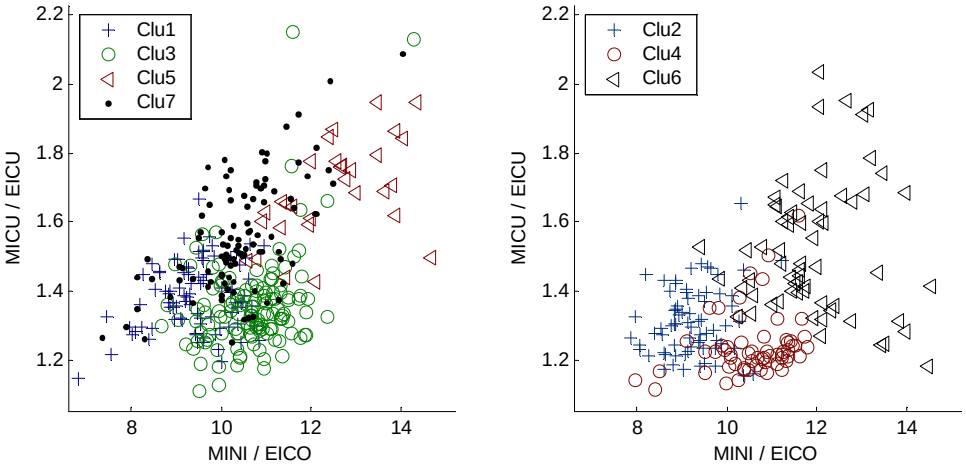


Figure 28. The best two-feature set to describe feed types

The mineralogical justification of the feature 'MICU / EICU' is clear: as the justification of JICU of Figure 27, the feature can be justified by the floatability of Copper. The other feature 'MINI / EICO' cannot be explained with my mineralogical know-how. Still, the feature contains information of the feed types, as it can separate the feed type clusters. Resolving the mineralogical background would add to my mineralogical understanding.

6.3.2 Authoring the knowledge base of the best practices

The other major task of the analysis of the Hitura case was to define the best process control practices for each feed type. This subsection studies the treatment of the most common feed type, the serpentinite, as defined by the cluster four of Figure 26. The first task was to decimate the data set to contain only the data points carrying the serpentinite label.

The second task was to re-label these data points according to the new goal of analysis: the success of process control. These labels are generated by clustering the success-SOM of Figure 29. This SOM was trained with the three variables: the SNI is the recovery of Nickel, the RNI is the Nickel content of the concentrate, and the RTON is the mass of the product. High SNI states that most of the valuables have been collected to the concentrate, high RNI correlates with a high quality of the product, and high RTON equals high production rate.

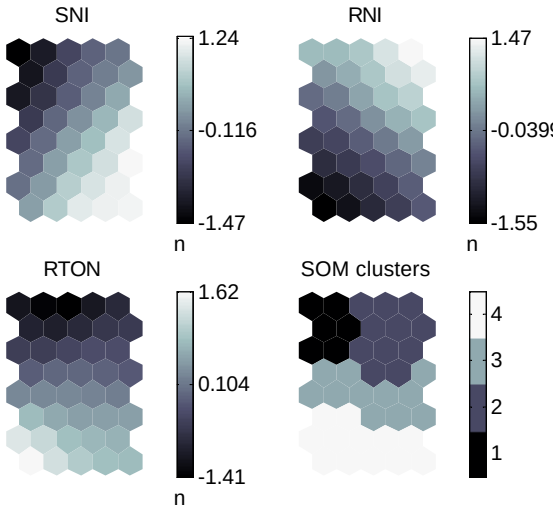


Figure 29. The SOM describing various approaches of refining the talc-amphibole feed type

The analysis is based on clusters one, two and four. The cluster three is ignored as it describes the average result. The cluster one depicts a poor strategy: product quality (RNI) is average, but both Nickel recovery (SNI) and production rate (RTON) are low. The cluster two depicts a strategy that emphasises product quality (high RNI). The cluster four emphasises high recovery (high SNI) and high production rate (high RTON), with the penalty of low product quality (low RNI). Of these three strategies, the strategies of cluster two and four are viable; the process operator should choose one of them depending on the requirements of the buyer. This concludes the problem definition phase; the remainder of this section studies, how to run the process to achieve the various strategies.

The studied on-line data comprised 55 on-line variables, mostly process controller set points. These points were labelled according to the SOM of Figure 29. The optimiser was requested to return the best pair of features; the result is in Figure 30. The feature ‘TM - KM2’ is the difference between the power draw of the rod mill (TM) and of the ball mill 2 (KM2); the feature ‘TM / KM1’ is the ratio between TM and power draw of ball mill 1 (KM1). The result suggests that to gain the process control strategy 2 (clu_2 in Figure 30) the operator should maintain high values of ‘TM / KM1’ and ‘TM - KM2’.

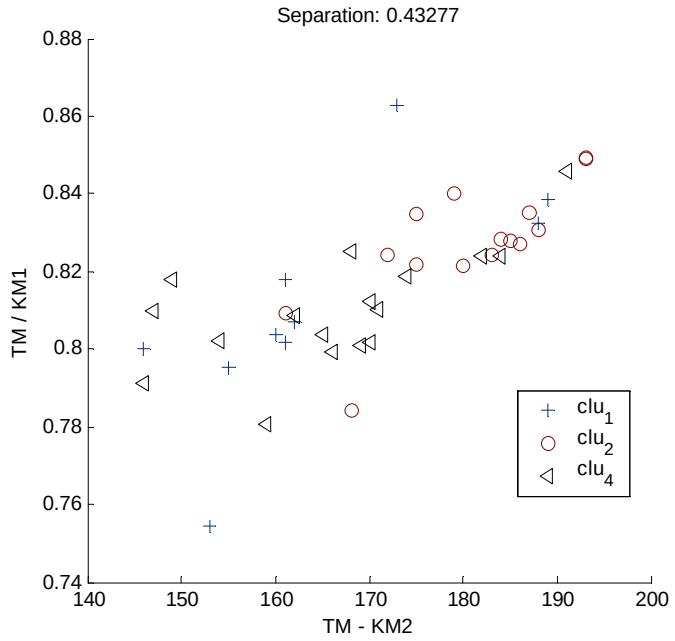


Figure 30. Two best two-feature combinations to explain differences in process control success

The above analysis could not separate the clusters one and four. Another analysis was performed with a data set that contained only the representatives of these two clusters. The result is in Figure 31: as the clusters separate, the two found features can be used to control the process to achieve the selected strategy.

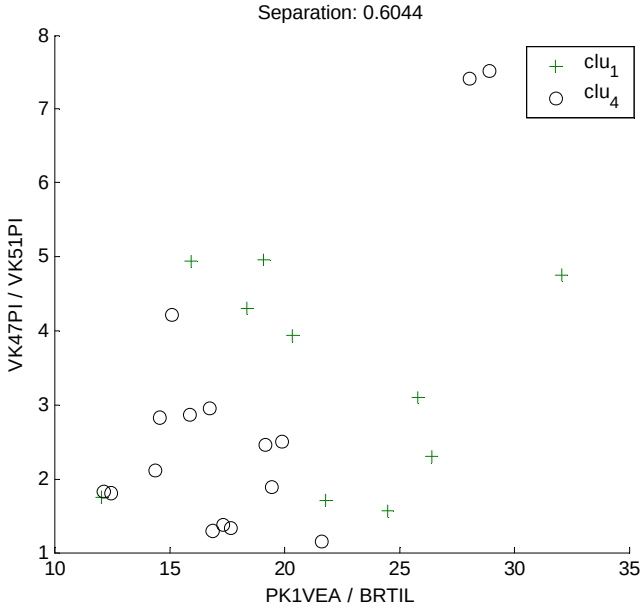


Figure 31. Two best features to explain the differences between process control strategies one and four

I cannot provide mineralogical justification to the above findings: I am not an expert of the control of the Hitura concentrator. However, because the success-clusters separate, the found features contain useful information. Resolving the mineralogical background would provide a concrete basis for the knowledge base of best practices.

6.4 Summary and discussion

Recall the three themes of Glymour et al. discussed in the subsection 2.3.2. The first theme was clarity about the goals of the analysis. This chapter showed that the goals can be clearly defined: in the first case, as the feed type of the concentrator; and in the second case as the strategy of treating a feed type. The second theme of Glymour was understandability of the strengths and weaknesses of the algorithm. The simplicity of the used methodology probably allows the user to understand, whether the approach of this thesis supports a given data analysis problem. The third theme was assessment of the reliability of the result. The simplicity of scatter-plot based presentation of results supports this task: if the clusters separate, the analyst result is reliable.

The results of this analysis have relevance to the practical process control at the Hitura mine. According to Pulkkinen et al (1993a) changes in the feed type are a problem. This was the basis of definition of the business goals (section 6.1) and study of the data (6.2). The solution to the problem was defined as an expert system that supports feed type based process analysis. The subsection 6.3.1 showed that a feed type estimator based on the on-line data of Courier 30 is feasible. And the subsection 6.3.2 showed, for the serpentinite feed type, that feasible process control practices can be learned from data.

7 Conclusions

According to my experience, process engineers seldom use statistics other than the simple operations of mean and variance. Still, as shown in chapter 6, these tools surface useful information from data acquired from the automation system. This is not surprising, as this data base is a rich collection of full-scale experiments, devised by process operators, and targeted to solve real problems.

This thesis claims that process engineers find statistical algorithms unreliable and difficult to understand. These issues are interrelated: if the user does not understand how to use a tool, he may misuse it. The main contribution of this thesis is to discuss the criteria that define proper tools for engineers, and to propose a methodology that complies with these criteria.

The goal of this thesis is to support learning from data. Thus, chapter 2 was devoted to the study of learning theories and practices. The core theory of this thesis is constructivism, which describes learning as construction of mental models. The goal of the methodology of this thesis is to support the construction of these hidden mental models. The next step was to bind this basic theory to the theories of organisational learning of Senge et al. Their work describes how the mental models influence the behaviour of a person, and how a group develops shared mental models. The next studied level was statistical methodologies designed to support the study of industrial data. This thesis showed two basic methodologies that match with the theories of Senge. This is not surprising, as both levels of theory describe the process of learning. However, this notion is useful, as it justifies the use of statistics to support organisational learning.

The above abstract study prepared grounds for the analysis of desirable properties algorithms intended for the process operator. The first criterion is supervised operation, which allows the user to control the process of browsing the data; the second is robustness, which is important as industrial engineers do not have time or statistical resources to pre-treat the data for specialised tools; the third is understandability, which may be the most important of these three criteria. Any analyst prefers understandable results. A basic tenet of this thesis is to create tools that are understandable to the process engineer.

With the above criteria, this thesis defined a methodology to support practical data analysis. In the first phase of this methodology, the user uses the SOM to perform exploratory data analysis; and to iterate this analysis until the SOM representation matches his understanding of the problem. This may require iterations in the selection of the variables and data points used to train the SOM; and in general, changes in the user's conception of the problem. The resulting formulation of the problem allows the proposed supervised methodology. The second phase comprises the use of variable selection and feature extraction to find information that is related to the user's problem definition; and the use of visualization to illustrate the results of search. The tools that perform this search need to be robust, as the user is bound to feed raw industrial data to the algorithm. The variable selection tool of this thesis is based on the k-NN type cluster separation measure, which makes no assumption of the shape of the clusters, and is robust towards outliers. The visualization of the results of the methodology is based on the scatter-plot, which understandably illustrates the manifestation of the phenomenon of interest in the data.

The above methodology respects the implicit nature of the mental models: the tools of this thesis are designed to allow an intuitive approach to data analysis. The first phase of the methodology of this thesis, the problem definition, is based on the SOM, whose training the user can iterate, until the SOM representation intuitively fits the user's understanding of the problem. A counter example is the Bayesian approach, in which the user is required to create a probabilistic model of his understanding of the problem. Also, the results of this thesis are presented on the intuitive level: the user is presented with scatter-plots that the user can study to further construct his mental models. The user is not forced to accept abstract constructs, such as rules, to his mental models. This may seem as a naïve approach to data analysis, but one should bear in mind that most process engineers are not fluent with statistics. Appreciation of intuitiveness is beneficial also because the goal of this thesis is to construct mental models in mind of the user, not onto the memory a computer.

The methodology of this thesis is illustrated using the case of the Hitura concentrator. The first application of the methodology was to study, how to create an on-line observer of the feed type. The problem was defined by a SOM mapping of the feed types of the concentrator, based on the data acquired from the off-line laboratory. The answer was defined by the on-line features that describe the feed type. After mineralogical justification, these features can serve as a basis for the on-line observer of the process feed type. The second application of the methodology was to study, how to find the best process control practices to treat each feed type, in this case, the serpentinite. The problem was defined by a SOM mapping of the chemical properties of the product, again based on off-line laboratory data. The answer was defined by the on-line features extracted that describe which process control parameters have highest leverage in the control of this feed type, and showed how they should be controlled to achieve each feed type. After mineralogical justification, these features can be deployed to serve process control. The main source of process control problems in the Hitura mine is the volatility of the feed type. With the methodology of this thesis, an engineer can create an on-line observer that alerts the user to changes in the process feed type; and the engineer can define a set of best practices that describe what to do after a feed type change. At Hitura, this is essential process control know-how.

The tools of this thesis can be used also in other domains than the concentrator of Hitura. In general, the methodology supports analyses targeted to study two data sets in parallel. The other data set is clustered, and the other data set is browsed for information that is related to this clustering. In Hitura, the first data set comprised laboratory data; the second data set comprised on-line data. Another study based on the methodology of this thesis is a study of emissions caused by various diesel fuels (Laine 2002). The first set of data was the analysis of the emissions of the engine exhausts; the second set of data was the chemical analyses of the diesel fuels. The goal of the analysis is to support the development of low-emission fuels.

The tools of this thesis are not created to impress champions of statistics. The tools are created to understandably and robustly provide information that the engineer finds interesting. Such tools help engineers to construct mental models; and help experts to share their mental models with colleagues.

This is learning from data.

Appendix A. The Tabu-search

Tabu-search is a greedy algorithm of combinatorial optimisation (see e.g. Kreher 1999): it always tries to advance in the direction where the gradient of the cost function is the highest, as illustrated by the first subplot of Figure 32. The special property of the Tabu-search is its tail, which forbids re-entry to already visited directions. As shown in the second subplot, this tail may fill up a local minimum, and force the search to continue towards new areas, possibly towards the global optimum, as illustrated in the third subplot.

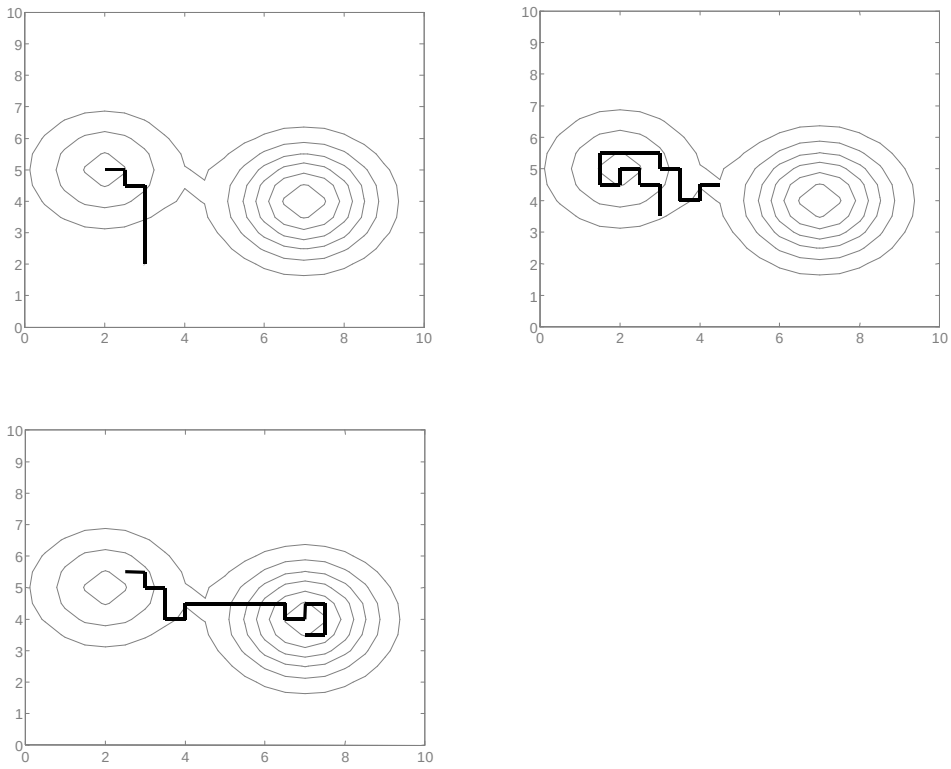


Figure 32. The route of the Tabu-search in the space of two maxima

Bibliography

Aha, D., Bankert, R., 1996, A comparative evaluation of sequential feature selection algorithms. In D. Fisher and H.-J. Lenz, editors, Learning from Data, chapter 4, pp. 199-206. Springer.

Akaike,H., Parzen, E., Tanabe, K., Kitagawa, G, 1998, Selected papers of Hirotugu Akaike, Springer series in statistics, Heidelberg.

Back, A., Tappenberg, T., 2001, Selecting Inputs for Modeling Using Normalized Higher Order Statistics and Independent Component Analysis, Ieee Transactions On Neural Networks, 3 (12) 612-617.

Bearman, R., Milne, R.W., 1992, Expert Systems: opportunities in the minerals industry, Minerals Engineering, (5) 1307-1323.

Bellman, R., Zadeh, L., 1970, Decision making in a Fuzzy environment, Management science, 17 (4) 141-164.

Bishop, C., Svensen, M., & Williams, C., 1996, GTM: a principled alternative to the self-organizing map, Proc. Int. Conf. on Artificial Neural Networks, ICANN'96, Springer-Verlag, 164-170.

Bonnlander, B., Weigend, A., 1994, Selecting input variables using mutual information and nonparametric density estimation, Proc. of the 1994 Int. Symp. on Artificial Neural Networks ({ISANN'94), 42-50

Buntine, W. L., 1996, Graphical Models for discovering knowledge, in Influence Diagrams, Belief Nets and Decision Analysis, ed: Fayyad, U., Plateletsky-Shapiro, G., Smyth, P., Uthurusamy, R., AAAI Press, Menlo Park.

Burges, C., 1998, A tutorial on support vector machines for pattern recognition. In Data Mining and Knowledge Discovery. Kluwer Academic Publishers, Boston, (2).

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., Wirth, R., 2000, CRISP-DM 1.0 step-by-step data mining guide, Technical report, CRISM-DM consortium. *http://www.crisp-dm.org*.

Charniak, E., 1991, Bayesian Networks without Tears, AI Magazine, Winter issue, (12) 50-63.

Chen, M., Han J., Yu, P., 1996, Data mining: an overview from a database perspective, IEEE Transactions On Knowledge and Data Engineering, 6 (8) 866-883.

Chernoff, H., 1973, The use of faces to represent points in k-dimensional space graphically, Journal of American Statistical Association, (63) 361-368.

Dasarathy, B, 1991, Nearest Neighbor (NN) Norms: NN pattern classification techniques, Washington DC, IEEE computer society.

Dash, M., Liu, H., 1997, Feature selection for classification, International Journal of Intelligent Data Analysis, 1 (3) 131-156.

Dash, M., Liu, H., Yao, J., 1997, Dimensionality Reduction for Unsupervised Data, Proceedings of the Ninth IEEE International Conference on Tools with AI, IEEE Computer Society Press, 532-539.

Dempster, A., Laird, N., Rubin, D., 1977, Maximum-likelihood from incomplete data via the EM algorithm, Journal of Royal statistical society, B, (39) 1-38.

Devijver, P. A., Kittler, J., 1982, Pattern recognition: a statistical approach, Prentice/Hall international, London.

Devroye, L., Gyoyfi, L., Lugosi, G., 1996, A probabilistic theory of pattern recognition, Springer-Verlag, Berlin

Domingos, P., and Pazzani, M., 1996, Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier**.** Proceedings of the 13 th Int. Conference on Machine Learning, 105-112.

Easterby-Smith, M., 1997, Disciplines of the learning organization: contributions and critiques. Human Relations, (50) 1085-1113.

Fayyad, U., Haussler, D., Stolorz, P., 1996a, KDD for science data analysis: issues and examples, KDD-96, AIII press, *http://www.aig.jpl.nasa.gov/kdd96*.

Efron, B., Tibshirani, R., 1993, An Introduction to the Bootstrap, Chapman and Hall, New York.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996b, From Data Mining to Knowledge Discovery in Databases, AI Magazine, 17(3)37-54.

Ferreira, C., Gene Expression Programming in Problem Solving, WSC6 tutorial, 2001.

Ferri, F., Pudil, P., Hatef, M, Kittler, J., 1994, Comparative study of techniques for large-scale feature selection.

Fisher, J., Principe, J., 1998, A Methodology for Information Theoretic Feature Extraction, In Proc. of IEEE World Congress On Computational Intelligence, Anchorage, Alaska, May 4-9, 1712-1716.

Fisher, R., 1936, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, (7) part II, 179-188.

Flexer, A., 1999, On the use of self-organizing maps for clustering and visualization, in Principles of Data Mining and Knowledge Discovery, 80-88.

Friedman, J., Stuetzle, W., 1981, Projection Pursuit Regression, Journal of the American statistical association, 376 (76) 817-823.

Gaag, L., Renooij, S., Wiiteman, C., 2002, Probabilities for a probabilistic network: a case-study in oesophageal cancer, Artificial Intelligence in Medicine (2002) to appear.

George, E., McCulloch, R, 1993, Variable selection via Gibbs sampling, Journal of the American statistical association, 88 (423) 881-889.

Glymour, C., Madigan, D., Pregibon, D., Smyth, P., 1997, Statistical themes and lessons for data mining. Data Mining and Knowledge Discovery, 1 (1) 11-28.

Györfi, L., 2002, Principles of non-parametric learning, ed: Györfi, L., In series CISM international center for mechanical sciences, number 434, Springer, New York.

Haddawy, P., 1999, An Overview of Some Recent Developments in Bayesian Problem-Solving Techniques, AI Magazine, Summer 1999.

Hand, D., 1994, Deconstructing statistical questions, Journal of the Royal statistical society, Series A, 3 (157) 317-356.

Hätönen, K., Laine, S., Similä, T. 2003, Using the LogSig-function to integrate expert knowledge to Self-Organizing Map (SOM) based analysis, 2003 IEEE International Workshop on Soft Computing in Industrial Applications, in press.

Himberg, J., 1998, Enhancing SOM-based data visualization by linking different data projections, Ed.: Xu, L., Chan, L.W., King, I., Intelligent Data Engineering and Learning (IDEAL ‘98), Springer, 427-434.

Huber, P., Robust Statistics. John Wiley & Sons, 1981

Hyvärinen, A., Karhunen, J., Oja, E., 2001, Independent Component Analysis, John Wiley & Sons, New Jersey.

Jaimes, A., Chang, S-F, 2000, Automatic Selection of Visual Features and Classifiers, In proceedings of IS&T/SPIE Storage and Retrieval for Image and Video Databases VIII, January 2000, vol. 3972.

Jain, A., Duin, R., Mao, J., 2000, Statistical pattern recognition: a review, IEEE trans. on pattern analysis and machine intelligence, 1 (22) 4-37.

Jain, A., Murty, M., Flynn, P., 1999, Data clustering: a review, ACM Computing Surveys, 3 (31) 264-323.

John, G., Kohavi, R., Pfleger, K., 1994, Irrelevant features and the subset selection problem, Proceedings of the eleventh international conference on machine learning, Eds: Cohen, W., Hirsh, H., Morgan Kaufmann Publishers, 121-129.

John, G., Langley, P., 1995, Estimating continuous distributions in Bayesian classifiers, Proceedings of the eleventh conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers, 338-345.

Jordan, M. Gharamani, Z. Jaakkola, T. and Saul, L. 1997, An introduction to variational methods for graphical models. Learning in Graphical Models, Kluwer Academic

Kachigan. S., 1991, Multivariate statistical analysis, A conceptual introduction, 2nd ed., Radius Press, New York.

Kaski, S., 2002, Clustering based on conditional distributions in an auxiliary space, Neural Computing, 1 (14) 217-240.

Kaski, S., Sinkkonen, J., Peltonen, J., 2001, Bankruptcy analysis with self-organizing maps in learning metrics. IEEE Trans. Neural Networks 12, 936-947

Keim, D., 2002, Information visualization and Visual data mining, IEEE transactions on visualization and computer graphics, 1 (8) 1-8

Keim, D., Kriegel, H-P., 1996, Visualization techniques for mining large databases: a comparison, IEEE transactions on knowledge and data engineering, 8 (6) 923-938.

Kohavi, R., 1995, A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI 1995, 1137-1143.

Kohavi, R., John, G., 1997, Wrappers for feature subset selection, Artificial Intelligence Journal, Special Issue on Relevance, 97 (1-2) 273-324

Kohonen, T., 2001, Self-Organizing Maps (Series in Information Sciences, 3rd ed., Springer, Heidelberg, 2001.

Koller, D. and Sahami, M. 1996, Toward optimal feature selection. Machine Learning: ICML' 96, Bari, Italy, 284-292.

Kreher, D.L., Stinson, D.R., 1999. Combinatorial Algorithms: Generation Enumeration and Search. CRC, Boca Raton.

Laaksonen, J., Oja, E., 1996, Classification with learning K-Nearest Neighbors, ICNN'96, 1480-1483.

Lagus, K., Alhoniemi, E., Valpola, H., 2001, Independent Variable Group Analysis, In: proceedings of Artificial Neural Networks - ICANN 2001, Eds. G. Dorffner, H. Bischof, K. Hornik, Austria, 203-210

Laine, S., 1997, Ore type based expert system for the Hitura concentrator, Licentiate thesis, Helsinki University of Technology, Espoo, Finland.

Laine, S., 2002, Early result of emission-analysis, internal report, Helsinki University of Technology, Espoo, Finland.

Lampinen, J., Vehtari, A., 2001, Bayesian Approach for Neural Networks – Review and Case Studies, Neural Networks, 14 (3) 7-24.

Langley, P., 1994, Selection of relevant features in machine learning. In AAAI Fall Symposium on Relevance, 140-144.

Lappalainen, H., Honkela, A., 2000, Bayesian nonlinear independent component analysis by multi-layer perceptrons. In M. Girolami, editor, Advances in Independent Component Analysis, Springer-Verlag, 93-121.

Lerner, B., Guterman, H., Aladjem, M., Dinstein, I., 1999, A Comparative Study of Neural Network Based Feature Extraction Paradigms, Pattern Recognition Letters, 1 (20) 7-14,

Li, C., Biswas, G., 2002, Unsupervised learning with mixed numeric and nominal data, Ieee trans. on knowledge and data engineering, 14 (4) 673-690.

Lindner, G., Studer, R., 1999, AST: Support for algorithm selection with a CBR approach. In C. Giraud-Carrier and B. Pfahringer, editors, Proceedings of the ICML-99 Workshop on Recent Advances in Meta-Learning and Future Work, Bled, Slovenia, 38-47.

Liu, H., Setiono, R., Feature Selection via Discretization, IEE trans. on knowledge and data engineering, 9 (4) 642-645.

Lowe, D. and Tipping, M. E., 1996, Feed-forward neural networks and topographic mappings for exploratory data analysis, Neural Computing and Applications 4, 83-95.

Madigan, D. and Rafferty, A., 1994, Model selection and accounting for model uncertainty in graphical models using Occam's window. Journal of the American Statistical Association, 89 (428) 1535-1546.

Mao, J., Jain, A., 1995, Artificial neural networks for feature extraction and multivariate data projection, IEEE Transactions on Neural Networks, 2 (6) 296 –317

Miller, 2002, A. J., Subset selection in regression, New York, Chapman-Hall.

Montanari, A., Lizzani, L., 2001, A projection pursuit approach to variable selection, Computational Statistics & Data Analysis 35 (2001) 463-473.

Murthy, S., 1997, Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, Kluwer academic publishers.

Parzen, E., 1962, On Estimation of a Probability Density Function and Mode, Annals of Mathematical Statistics, (33) 1065-1076.

Pyle, D., 1999, Data preparation for data mining, Morgan Kaufmann Publishers, Kent.

Phillips, D., Soltis, J., 1998, Perspectives on Learning, 3rd ed, Teachers college press, New York.

Ray S., Chan., A., 2001, Automatic feature extraction from wavelet coefficients using genetic algorithms, Neural Networks for Signal Processing XI. Proceedings of the 2001 IEEE Signal Processing Society Workshop, 233 –241.

Ribeiro, R., 1996, Fuzzy multiple attribute decision making: a review and new preference elicitation techniques, Fuzzy sets and systems, 78 (2) 155-181.

Sammon, J., 1969, A non-linear mapping for data structure analysis, IEEE transactions on computers, 5 (C-18) 401-409.

Schölkopf, B., Smola, A., Muller, K., 1998, Nonlinear Component Analysis as a Kernel Eigenvalue Problem,º Neural Computation, 10 (5) 1299-1319, .

Schwarz, G., 1978, Estimating the dimensions of a model, Annals of Statistics, (6) 461-464

Senge, P., 1994, The Fifth Discipline, Currency/Doubleday.

Senge, P., Kleiner, A., Roberts, C., Ross, R., Smith, B., 1994, The fifth discipline fieldbook, Nicholas Brealey Publishing.

Siponen, M., Vesanto, J., Simula, O., Vasara, P., 2001, An Approach to Automated Interpretation of SOM. In Proceedings of Workshop on Self-Organizing Map 2001 (WSOM2001), Springer, 89–94.

Skalak, D., 1994, Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms. In The 11th International Conference on Machine Learning, 293-301.

Smith, M. and Kohn, R., 1996, Nonparametric regression using Bayesian variable selection. J. Econometrics, (75) 317-344.

Svensén, M., 1998, Generative Topographic Mapping (GTM), Doctoral Thesis, Aston University, Birmingham, UK. *http://www.ncrg.aston.ac.uk/GTM/*

Swiniarski, R., Skowron, A., 2003, Rough set methods in feature selection and recognition, Pattern recognition letters, (24) 833-849.

Torkkola, K., 2001, Nonlinear feature transforms using maximum mutual information, in Proceedings of the IJCNN 01, 2756-2761.

Torkkola, K., 2002, On feature extraction by mutual information maximization. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (1) 821-824.

Torkkola, K., Campbell, W., 2000, Mutual information in learning feature transformations. In Proceedings of the 17th International Conference on Machine Learning, Stanford, CA, USA, June 29 - July 2, 1015-1022.

Tufte, E., Envisioning information, Cheshire, Conn.: Graphics press, 1990.

Tufte, E., The visual display of quantitative information, Cheshire, Conn.: Graphics press, 1983.

Ultsch, A., 1993, Self organized feature maps for monitoring and knowledge acquisition of a chemical process. In S. Gielen and B. Kappen, editors, Proceedings of the International Conference on Artificial Neural Networks, Springer-Verlag, 864–867.

Vafaie, H. and De Jong, K.A., 1993, Robust feature selection algorithms, In Proceedings of the International Conference on Tools with AI, Boston, MA, IEEE Computer Society Press, 356-364.

Vapnik, V., Statistical Learning Theory. New York: John Wiley & Sons, 1998.

Vesanto, J., 1999, SOM-Based Data Visualization Methods, Intelligent Data Analysis, 2 (3) 111-126

Vesanto, J., 2002, Data Exploration Process Based on the Self- Organizing Map, Doctoral thesis, Helsinki University of Technology, Finland. *http://lib.hut.fi/Diss/2002/isbn9512258978/*

Vesanto, J., Alhoniemi, E., "Clustering of the self-organizing map". IEEE Transactions on Neural Networks, 11(3) 586–600, 2000.

Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J., 1999, Self-organizing map in matlab: the SOM toolbox. In Proceedings of the Matlab DSP Conference, Espoo, Finland, 35-40. The toolbox is available from *http://www.cis.hut.fi/projects/somtoolbox/*.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V., 2000, Feature selection for SVMs. In Advances in Neural Information Processing Systems 13. MIT Press, 668-674.

Wheatley, M., 1999, Leadership and the new science, Berret-Koehler Publishers.

Yang, H., Moody, J., 1999, Data visualization and feature selection: new algorithms for nongaussian data, NIPS'99, vol. 12.

Yang, Y., Pedersen, J. O. 1997, A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, US, 1997), D. H. Fisher, Ed., Morgan Kaufmann Publishers, San Francisco, US, 412-420.