IDENTIFICATION OF DATA MINING TECHNIQUES FOR INDUSTRIAL PROCESS ANALYSIS AND CONTROL

Edith Namikka and George J. Gibbon

School of Electrical and Information Engineering,
University of the Witwatersrand, Johannesburg,
Private Bag 3, WITS, 2050.
Email: e.namikka@ee.wits.ac.za, g.gibbon@ee.wits.ac.za

Abstract: This paper describes data mining techniques for application to industrial process analysis tasks. The aim was to identify data mining techniques that support exploratory analysis for performance assessment, process modelling, and fault diagnosis - tasks that form the foundations of process control. The data mining technique base was investigated by conducting experiments using both synthetic and real process data. The assembled techniques included clustering methods, the multivariate statistical method of principal component analysis, and outlier detection methods. The selected techniques were implemented and tested using real process data, and the results demonstrated the usefulness of the methods for process performance analysis, modelling for historical analysis, and fault detection and diagnosis. *Copyright © 2002 IFAC*

Keywords: Data processing, modelling, statistical process control, fault detection.

1 INTRODUCTION

Industrial processes are invariably large and complex, and generate large amounts of high dimensional data. This data contains potentially useful information that can be invaluable in the control and optimisation of processes, if the appropriate tools are applied to the data. Industrial process analysts are thus increasingly looking to data treatment methods to aid in the analysis of process performance for purposes of improving process efficiency, productivity, and product quality. This paper investigates data mining methods that are useful for the analysis of industrial process data.

1.1 Background

Data mining is a wide discipline concerned with the extraction of implicit, and potentially useful information from data, by searching for patterns and relationships hidden in the data (Zaiane, 2002). There are five major categories of data mining methods, classified according to the type of knowledge to be mined (Zaiane, 2002), they include:

i. Classification involves building a model for one attribute (variable) – the class attribute, as a function of other variables. The class attribute is usually quantitative and discreet in nature.

ii. Deviation detection involves discovering and analysing significant changes in data, compared to previously measured or normative values.

iii. Clustering aims to find groupings (clusters) in data such that points in the same cluster are similar to one another and points in separate clusters are dissimilar.

iv. Association rule discovery produces dependency rules which predict the occurrence of one item based on the occurrence of other items.

v. Sequential pattern discovery finds rules that predict strong sequential dependencies among events (Zaiane, 2002).

Data mining methods are drawn from different fields including statistics, machine learning, database systems, and pattern recognition (Reinartz, 1999). Another category of data analysis methods is the field of multivariate statistical methods. This field of techniques has to be considered in any work that is concerned with industrial process analysis. The reasons for this will be discussed in later sections.

Data Mining in Industrial Process Analysis and Control Applications: Although today's plant systems are very effective in data collection and storage, the advances in data gathering, storage, and distribution technologies have far outpaced the advances in data analysis techniques for industrial applications (Agrawal et al., 1996). Increasingly, process engineers have looked to data mining for methods to handle process data (Tsai et al., 1986). The importance of data methods in process control-related applications is appreciated by considering the fact that the development of any process monitoring or control strategy, requires the following steps;

i. attaining a clear understanding of the process,
ii. modelling the process,
iii. designing the control or monitoring solution, and
iv. implementing the solution.

The bulk of the work is usually in steps i and ii which are usually non-trivial because of the typically complex nature of industrial processes. By allowing the extraction of relationships in data, data mining approaches have proven suitable for the analysis of large volumes of data, aiding analysts to understand and model processes for which physical models are difficult or impossible to develop (Tsai et al., 1986). The process-related applications in which data mining techniques have been used can generally be grouped into five categories;

i. process analysis for performance assessment,
ii. process modelling,
iii. prediction,
iv. anomaly detection and diagnosis, and
v. process control and optimisation.

1.2 Objectives and Scope of the Project

The pool of techniques in the data mining discipline is very large, so a study describing the major data mining functionality required for industrial process analysis applications is essential if a process-control data mining system were to be implemented. (This project[1] was in fact the foundational work for the development of a process control-data mining tool). As data mining is a well-researched discipline, the aim of this project was not to develop new data mining techniques, but to identify suitable methods and adapt them for application to industrial process analysis-related tasks that are considered the areas of significance in process analysis and control system development, namely:

i. exploratory analysis and process examination for performance assessment,
ii. process modelling (for inferential purposes, monitoring, or control development), and
iii. anomaly detection.

The data mining methods assembled to meet the requirements of industrial process analysis were grouped into three categories: clustering methods, multivariate statistical methods, and outlier detection methods. Four clustering algorithms were identified: a hybrid algorithm combining k-means and subtractive clustering, fuzzy c-means, basic minimum-squared-error, and subtractive clustering. The multivariate statistical method used was principal component analysis (PCA), and the outlier detection methods included a distance-based method, and a local-density-based method. Reasons for the selection of each method, as well as the functionality requirement met by each method are described in later sections.

Section 2 of the paper presents the methodology used in the technique identification, and section 3 describes the selection of algorithms. The experiments demonstrating the efficacy of the methods are only briefly discussed in this paper, and readers are referred to a complimentary paper (Namikka, 2003) for the detailed experimental work. Sections 4 and 5 conclude the outcome of the project.

2 METHODOLOGY OF THE TECHNIQUE SELECTION PROCESS

Two criteria were used in investigating potential data mining methods for industrial process analysis, namely:

i. The required functionality for process-related analyses, as described in section 1.2 above.

ii. The nature of industrial process data, particularly the two ubiquitous characteristics of process data - high dimensionality, and inconsistencies such as missing values, noise and outliers. The other factor that drove the investigation was the computational characteristics (such as speed) associated with a given algorithm.

By considering the typical applications of the different data mining categories, an initial high-level selection of the categories that were considered useful in industrial process analysis was undertaken. Table 1 lists the typical applications of the different data mining categories. Two categories that were eliminated at this initial stage were association rules and sequential pattern discovery. This was because typical applications of these two categories have been in non-industrial applications (Table 1). Although classification methods have been applied to some process tasks such as process characterisation (Aldrich and Schmitz, 1997), the category was not used in this project because of the following reasons:

i. Classification mainly deals with categorisation into discrete 'classes', say in the identification of defective parts. These types of applications are normally used in discrete manufacturing.

ii. Classification involves "supervised" clustering, where the number and types of groups into which data is to be grouped is already established. Yet with industrial process data, the inherent grouping of a data set is seldom known, let alone the number of groups.

This work therefore focused on clustering methods, deviation detection methods, and multivariate statistical methods. A critical review of the literature on the selected technique categories was undertaken to identify an initial list of suitable algorithms. The 'short-listed' algorithms were then compared by running tests using both synthetic and real data, before the final selection was made.

[1] This project is sponsored by OPTI-NUM Solutions, MATLAB's partner in Southern Africa.

Table 1: Typical applications of different data mining technique categories (Zaiane, 2002).

	Data Mining Category	Data Mining Function	Typical Applications
1	Clustering	Unsupervised classification	Exploratory data analysis, image processing, dimension reduction, modelling for prediction, control development.
2	Classification	Supervised classification	Market analysis, fraud investigation, medical diagnosis, defective parts identification, process characterisation.
3	Deviation/Anomaly Detection	Outlier analysis	Fraud investigation, fault diagnosis, defective parts checks.
4	Association Rules Discovery	Association	Business data analysis/decision support (market analysis, resource planning, pricing analysis).
5	Sequential Pattern Discovery	Trend analysis	Business data analysis/decision support (risk analysis and management, business forecasting), medical applications, spatial and geographic applications.
6	Multivariate statistical methods	Lower dimensional projection	Dimension reduction, process modelling for monitoring or assessment, anomaly detection, statistical process control.

3 IDENTIFICATION OF SUITABLE DATA MINING METHODS

3.1 *Clustering Methods*

There are a myriad of clustering algorithms, with most falling under one of two main categories namely hierarchical and partitional algorithms (Jain et al., 1999). Other types of clustering algorithms include density-based methods, evolutionary algorithms, neural networks, and clumping methods. A full survey of the different clustering approaches is not discussed here, and readers are referred to (Jain et al., 1999, He, 1999). For clustering in process analysis applications, the following were the factors that the algorithms were required to handle: (i) high dimensionality, and (ii) inconsistencies such as missing values, outliers, or noise. Based on the above considerations, six clustering algorithms were identified for initial investigation; two hierarchical algorithms namely the nearest-neighbour (NN) and farthest-neighbour (FN) algorithms; three partitional algorithms, namely the k-means, basic iterative minimum-squared error (basic MSE), and fuzzy c-means (FCM) algorithms, and a density-based algorithm known as subtractive algorithm. The details of each algorithm are not discussed in the paper, and only the salient characteristics of each are described for purposes of a comparative analysis.

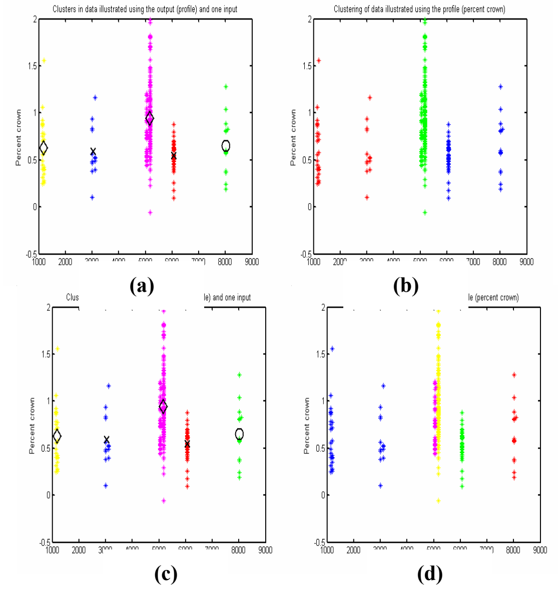
The FN and NN are both agglomerative hierarchical techniques, NN links the closest points of the clusters being merged, whereas FN merges clusters whose farthest points are closest (Jain et al., 1999). The basic MSE and k-means algorithms both form clusters that minimize the sum-of-squared-error function. The difference between them is that basic MSE is a sequential process, whereas k-means is a batch process (Jain et al., 1999). FCM does not generate distinct partitions but forms fuzzy clusters, allowing data points to have membership in all the clusters (Bezdek in Jang et al., 1997). Each of the selected algorithms was implemented using MATLAB, and the characteristics of each algorithm studied using firstly synthetic data, and then real process data. The synthetic data sets A, B, and C that were used consisted of the following: data set A was a random selection of data points with no specific groupings, B contained three compact and linearly separable clusters, and C comprised of two linearly separable clusters of unequal size. All the algorithms work well and produce the same results when the data set contains compact, linearly separable clusters of roughly equal sizes (data set B). But when the clusters are disparate, very different results are produced by each method. Table 2 summarises the findings of the investigation using synthetic data.

Table 2: Results from the analysis with synthetic data - characteristics of the different clustering algorithms.

Algorithm	Strengths or Weaknesses
Basic MSE	▪ When one of two well-separated and close clusters is much larger than the other, it splits the larger cluster to produce two clusters of roughly the same size (data set C). ▪ Sensitive to the order in which the feature vectors are presented. Different results are achieved for different permutations of the data set.
K-Means	▪ Dependent on the choice of initial cluster centres. ▪ Also splits a larger cluster to produce two equal ones when one of two well-separated and close clusters is much larger than the other (data set C).
Farthest-neighbour	▪ Discourages the production of elongated clusters (Jain et al., 1999). Produces meaningless groupings when clusters are not compact and/or well separated.
Nearest-neighbour	▪ Highly sensitive to the details of the data, and tends to form elongated clusters - an effect known as the “chain effect” (Jain et al., 1999).
Fuzzy C-means	▪ Not dependant on the choice of initial clusters, or order of data points, same results are obtained each time.

The sensitivity of the k-means algorithm to the choice of initial cluster centres (due to convergence to local minima) is undesirable, as it would require accurate choices of the initial centres. To deal with the problem, another clustering method known as subtractive clustering was combined with k-means to form a hybrid algorithm. Subtractive clustering is density-based and automatically searches for cluster centres based on a measure of density (Chiu in Jang et al., 1997). Subtractive clustering does not require inputs of the initial cluster centres or cluster number.

The Hybrid Clustering Algorithm: The hybrid algorithm was developed to combat the shortcoming of the k-means method, while retaining the desirable qualities of the method. K-means is desirable because of its simplicity and high efficiency - the algorithm has a complexity of $O(nk)$ where n is the number of data items and k the number of clusters. K-means is also order-independent. The subtractive clustering algorithm finds regions in the data with high densities of points, and the point with the highest number of neighbours is selected as centre for a cluster. The process is then repeated (Chiu in Jang et al., 1997). The hybrid algorithm uses subtractive clustering to first generate the optimal cluster centres, then these centres are fed to k-means as the initial cluster centroids. This hybrid algorithm was proven to be superior to the k-means algorithm, as the hybrid accurately produced the same results for each clustering process; this is demonstrated in Figure 2. A real process data set (industrial data from an aluminium rolling mill) was used to test the hybrid algorithm; the data contains five natural clusters corresponding to five aluminium alloys.



(a) (b) (c) (d)

Figure 1: Superiority of hybrid algorithm; the hybrid algorithm accurately produces five clusters in each run - (a), (c), while k-means produces different clusters (number and content) each time – (b), (d). (Same colour points belong to the same cluster).

The next step was to test the algorithms on real process data; each of NN, FN, FCM, basic MSE and the hybrid algorithm were used to cluster the aluminium profile data. The data consisted of 270 items with five variables. NN and FN were found to be too sensitive to the details of a data set (produced meaningless clusters which do not capture data groupings, due to the algorithms' tendencies described in Table 2.) Also, the clustering process with NN and FN was too slow even for a data set of modest size such as the one used in these tests (NN and FN took 14.2 min and 16.42 min respectively, compared to 0.48 s, 18.48 s and 1.22 s for the hybrid, basic MSE and FCM respectively) (Namikka, 2003). Hence FN and NN were eliminated and the following four were selected as the most suitable clustering algorithms for industrial data analysis: the subtractive – k-means hybrid algorithm, FCM, basic MSE, and subtractive clustering.

Other tests were done to test the efficacy of the selected in process analysis, the tests included:

- Exploratory analysis tests using another data set from the aluminium rolling mill. The data consisted of measurements of roll and bend forces, temperature, and the sheet profile measurements
- Fuzzy modeling was done using data from a tunnel-temperature measurement experiment that was designed specifically for this project. Temperature conditions in the tunnel were varied by changing the amount of air let through the tunnel and the heat setting on a blower. The aim was to mimic conditions in any typical plant in which temperature changes occur due to say chemical changes. Each data set consisted of 1500 records, and eight variables. The FCM and subtractive clustering methods were used in the modelling experiments, and the algorithms were proven invaluable in fuzzy modelling process.
- Prediction: the fuzzy models generated were used to predict the values of a sensor's output during both normal operating conditions, and a period of operation with a faulty sensor. The fuzzy models were demonstrated to be good prediction tools. The details and results of the experiments were treated in the complimentary paper (Namikka, 2003).

3.2 *Multivariate Statistical Methods*

Multivariate statistical methods are suited to the typically high-dimensional nature of industrial data because they can be used to analyse relationships between several variables simultaneously. The multivariate statistical techniques that are of interest in process control are the multivariate projection methods, which constitute principal component analysis (PCA) and partial least squares (PLS). The PCA and PLS methods are both based on eigenvector decompositions, they essentially allow the projection of data of high dimensionality onto a lower dimensional space by representing large numbers of correlated process variables using a small number of latent variables (Wise 1991; MacGregor et al., 1991). The process data can then be examined in the lower dimension spaces to analyse process behaviour. This lower dimensional projection is particularly effective with process data because, although massive amounts of data are collected, variables are not all independent of one another because only a few underlying events drive all the variables.

In selecting the multivariate statistical technique(s) to implement in this project, the basic requirements for analysis of process data (ability to handle large numbers of variables and data inconsistencies) were again considered. PCA and PLS have both been widely documented by researchers and proven to have the ability to handle both of the process data problems mentioned. Because both methods are complex in nature, rapid implementations of each method could not be done for purposes of a comparative study (as was done for clustering techniques), rather, the two methods were investigated through a critical review of literature. The following were the main findings:

i. PCA and PLS work on the same principles, that is, both methods are used to obtain lower dimensional projections of data, which are then used for further analyses. MacGregor (1991) and Wise (1991) state that PLS can be used in exactly the same way as PCA for the monitoring of processes, historical analysis, and anomaly detection.

ii. PCA does not need assumptions on the underlying distribution of the data (Joliffe, 1986), this is ideal for process data analysis because the distribution of a process data set is seldom known.

Based on the above two fundamental features of the methods, PCA was selected as the ideal technique for this work. The method is described briefly (Joliffe, 1986). In PCA, an *m* by *n* data matrix **X** is decomposed into the sum of the product of *n* pairs of vectors, where each pair consists of a vector in *n* called the principal component (PC) loadings, and a vector in *m* called the scores. Generally it is found that the data can be adequately described using far fewer PC's than original variables, and the cross-validation method (Wold, 1978) is deemed the most reliable method of determining the optimal number of PC's to be retained. Once the optimal number of PC's has been determined, the original data is now approximated by a lower dimensional model which provides the basis for projecting (and analysing) any new data. Two statistics are commonly employed in using PCA: a *lack of model fit* statistic, Q, and a measure of the *variation within* the PCA model - Hotelling's T^2 statistic. Control limits are developed Q and T^2, so that unusual events in the data can be detected from data points that fall outside the limits.

Several tests were done to demonstrate the efficacy of the technique for process performance checking, and fault diagnosis. Only the salient features of each experiment are described here, the details and results were treated in Namikka, 2003. The tests, which were done using data sets from the tunnel-temperature measurement experiment include:

- Modeling for historical performance analysis: a PCA model of the process was built using data from a period of normal operation. Q and T^2 control limits (99th percentile) on test data showed the model to be a good representation of the system.
- Fault detection: Tests were done by applying the PCA model to three data sets from periods of faulty operation (different faults in each data set). Results showed the T^2 and Q statistic values going above the control limit, which indicates that something is going on that was not in the original data – hence a fault was detected each time.
- Fault diagnosis: To determine the cause of the fault in each test, contribution plots were used to examine the contribution of each variable to the Q statistic. The contribution plots always indicated the variable responsible for the unusually large Q. The PCA method was hence demonstrated to be useful in not just detecting faults, but also in establishing the cause of the fault. The tests showed that the PCA technique allows for the construction of linear models which are effective for performance and fault analysis, because the models adequately describe process fluctuations around an operating point.

3.3 Outlier Detection Techniques

While most categories of data mining are focused on the large percentage of objects in a data set, methods in the outlier detection category, focus on the tiny percentage of data points – outliers, that are often discarded before most analyses. These outliers – data points that are grossly dissimilar from the rest of the data, may be the subject of interest when rare events are a bigger concern than commonly occurring trends (Breunig et al., 2000). Outlier detection methods have been proven useful in the detection of abnormal plant behaviour and defective instrumentation (Knorr et al., 1998). Researchers have developed several outlier detection methods, most of which define outliers based on proximity measures relative to the rest of the data. There are five general approaches used in outlier detection as summarised in Table 3.

From the analysis of each outlier detection approach (Table 3), an initial selection of the suitable approaches for application to industrial data was done as follows;

i. Univariate statistical methods were unsuitable because the distribution of a process data set is typically unknown and difficult to determine.

ii. Depth-based methods could not be used as process data is typically of high dimensionality.

iii. Clustering methods were undesirable for the reasons described in Table 3.

The multi- dimensional distance-based and the local-density-based methods were then selected as the outlier detection techniques suitable for process data. The main features of the techniques are described.

The multi-dimensional distance-based method (DB), first developed by Knorr et al. (1998), was implemented using a nested-loop algorithm that allows for ranking of the outliers. The definition of DB outliers is based on the k^{th} nearest neighbour of a point (Ramaswamy et al., 2000). For a *k* and point *p*, $D^k(p)$ denotes the distance of the k^{th} nearest neighbour of *p*. Intuitively, points with larger $D^k(p)$ values have sparse neighbourhoods and are more likely to be outliers than points in dense neighbourhoods with smaller $D^k(p)$ values. A user specifies the suspected number of outliers *n*, which are then defined as follows: a point *p* is an outlier if no more than *n*-1 other points in the data have a higher $D^k(p)$ value than *p*.

Table 3: Comparison of Outlier Detection Methods

Method	Gist of Method	Strengths/Weaknesses
Univariate Statistical Methods (Barnett and Lewis, 1994)	Data is modelled using a stochastic method, then outliers are determined depending on their fit with this model.	• Methods are typically univariate. • Based on the assumption that the distribution of the data set is known or can be determined.
Multi-dimensional Distance-based (Knorr et al., 1998)	Outliers are detected based on the full dimensional distance of points from each other.	• Have the most computationally tenable algorithms for data of high dimensionality. • Allow for ranking of "outlier-ness".
Local-density-based (Breunig et al., 2000)	The "outlier-ness of a point is determined by analysing the local density of the point's neighbourhood.	• Allow points to have "degrees" of outlier-ness, and not a binary value. • Cater for disparate-density regions in a data set.
Depth-based (Preparata and Shamos, 1988)	Each data point is represented as a point in an n dimensional (n-D) space and is assigned a depth. Outliers are then likely to be those points contained in smaller depths.	• Inefficient for high-dimensional data sets ($n \geq 4$) because for an n-D data space, the analysis relies on the computation of n-D convex hulls, and the lower bound complexity of computing an n-D convex hull is $\Omega(N^{n/2})$ for N objects.
Clustering	Used to isolate outliers as a side-product (e.g. single points that do not fall into any of the generated clusters).	• Unreliable results because clustering algorithms are developed to optimise clustering, not outlier detection (Knorr, et al., 1996).

In the local-density based method (Breunig et al., 2000), outliers are determined relative to densities in their local neighbourhoods, and a parameter known as the local outlier factor (LOF) is used to indicate the degree of outlier-ness of a point. The outlier factor is local because only a restricted neighbourhood of each object is considered. Two parameters are used to define the notion of density: a parameter *MinPts* specifying a minimum number of objects, and a parameter specifying a volume. These two parameters determine a density *threshold,* so that regions are connected if their neighborhood densities exceed the given density threshold. The developers of this method demonstrated it to be useful in isolating outliers that would not be easily detected by approaches that treat outliers in a "global" sense.

The DB outlier detection method was implemented in MATLAB, and experimental tests done to test the algorithm's effectiveness using the aluminium sheet profile data. The tests were done as follows:

- Basic outlier detection: Six artificial outliers were inserted in the data set and a value of n = 6 used in the DB method. In the results, the abnormal points were always accurately detected as outliers.
- Outlier ranking: To illustrate the usefulness of the ranking mechanism of the DB method, the same profile data was used for another test but with the value of n set to 15 (even though only 6 outliers were present in the data as before). In the results, the 15 points with the highest values of $D^k(p)$ were isolated as outliers. However, from the $D^k(p)$ value of each point, it was clear that the additional 9 data points had very low $D^k(p)$ values compared to the true outliers. So the ranking in the method can be used to discard 'false' outliers. Hence the DB algorithm was proven an effective outlier detection technique for high dimensional data.

Outlier detection tests on the local-density based method were done using the same aluminium profile data, as well as the aluminium rolling process data. The LOF method was found to be useful when the data has clusters of different densities, so that using the same distance measure across the whole data set could render inaccurate results.

4 DISCUSSION OF RESULTS

The data mining methods identified (and proven) as suitable for industrial process analyses are summarised in Table 4.

From the experiments done to test the usefulness of the different selected techniques, three major outcomes were achieved from this work;

i. It was found that data mining methods are invaluable for process performance assessment (Namikka, 2003). The techniques that were shown to be effective for process analysis were clustering methods, and PCA. The clustering methods are suitable to explore process data for which the natural grouping of a data set is seldom known. The PCA method deals with the more challenging task of the simultaneous interpretation of multiple observations.

ii. The experimental work showed that process modelling can be done using data mining methods instead of the conventional modelling methods. PCA was successfully used to model for historical analysis or monitoring purposes, and the FCM and subtractive clustering algorithms were used in fuzzy-based modelling. The PCA and fuzzy modeling methods are able to represent non-linear relationships that are difficult to model with conventional methods. The methods are also robust under the influence of noise, and require no prior knowledge of data distributions.

iii. The last major finding of this work was that data mining techniques are vital for fault diagnosis, which is a foundational component of process control. The outlier detection methods and the PCA technique were both proven effective for this task.

Table 4: Data mining techniques for process performance analysis and control strategy development.

Category	Algorithm/Technique	Use in Process Analysis
Clustering	Hybrid (K-means/ subtractive clustering)	Exploratory analysis, performance assessment, visualisation of multivariate data, decision support.
	FCM	Fuzzy modelling (for process control and prediction), anomaly detection, dimension reduction.
	Basic MSE	Performance analysis, data visualisation, dimension reduction.
	Subtractive	Fuzzy modelling (for process control and prediction)
Multivariate Statistical Methods	PCA	Process modelling (for monitoring and inferential purposes), historical data analysis for performance assessment, fault detection and diagnosis, dimension reduction.
Outlier Detection	Distance-based	Fault detection and diagnosis.
	Local-density-based	Fault detection and diagnosis.

5 CONCLUSION

The aim of this project was to identify data mining techniques for industrial process applications. Methods were required to facilitate typical tasks in process-related analyses, namely process performance analysis, process modelling, and anomaly detection. The suitable data mining techniques were identified through critical literature reviews of data mining methods, and tests on different algorithms using synthetic and real process data. The techniques devised for the work are summarised in Table 4. Each technique was implemented and experimental tests done with data sets from a hot rolling mill, and from a temperature measurement experiment as described in section 3. The outcomes of this project were demonstrations of the usefulness of data mining methods for process performance analysis, process modelling for prediction, monitoring, or historical analysis, and fault detection and diagnosis. The methods essentially form the foundation for the development of any control strategy, as process modelling and fault diagnosis form vital parts of control systems.

REFERENCES

Agrawal, R., A. Arning, T. Bollinger, M. Mehta, J. Shafer and R. Srikant (1996). The Quest Data Mining System, *Proc. of 2nd International Conf. on KDD and Data Mining*, Portland, Oregon.

Aldrich C. and G.P.J. Schmitz (1997). Characterisation of Process Systems with Decision Trees Extracted from Neural Networks Models. *EUFIT*, Sept. 1997.

Barnett V. and T. Lewis (1994). *Outliers in Statistical Data*. John Wiley & Sons, NY.

Breunig M.M., H-P. Kriegel, R.T. Ng and J. Sander (2000). LOF: Identifying Density-Based Local Outliers. *Proc. of ACM SIGMOD Int. Conf. On Management of Data*, Dallas, Texas, 2000.

He Qin (1999). A Review of Clustering Algorithms as Applied in IR. *Thesis*, Graduate School of Library and Information Science, University of Illinois, 1999.

Jain, A.K., M.N. Murty and P.J. Flyn (1999). Data Clustering: A Review. *ACM Computing Surveys*, 31(3).

Jang, J-S.R., C-T. Sun and E. Mizutani (1997). Neuro-fuzzy and Soft Computing, number isbn 0-13-261066-3 in *Matlab Curriculum Series*, Prentice Hall, Upper Saddle River, NJ, USA.

Jantzen, J. (1998). Neurofuzzy Modeling, *Tech. Report no 98-H-874 (nfmod)*. Technical University of Denmark, Department of Automation, Lyngby, Denmark.

Joliffe L. T. (1986). *Principal Components Analysis*. Springer-Verlag.

Knorr, E., R.T. Ng and V. Tucakov (1998). Algorithms for Mining Distance-based Outliers in Large Datasets. *Proc. of VLDB Conference*, Sept. 1998.

MacGregor, J.F. and T. Kourti (1995). Statistical Process Control of Multivariate Processes. *Control Engineering Practice*, **3(3)**, 403-414.

MacGregor J.F., T.E. Marline, J. Kresta and B. Skagerberg (1991). Multivariate Statistical Methods in Process Analysis and Control. *AIChE Symposium Proc. of 4th Int. Conf. On Chemical Process Control*, Padre Island, Texas, February 1991, **P-67**, 17-22.

Namikka E. (2003). Data Mining for Process Performance Analysis and Control System Development. *MSc. Thesis*, University of the Witwatersrand, Johannesburg, 2003.

Preparata F. and M. Shamos (1988). *Computational Geometry: An Introduction*. Springer-Verlag, New York Inc., 1988.

Ramaswamy S., R. Rastogi and S. Kyuseok (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. *Proc. of ACM SIGMOD Conf.*, 2000, 427-438.

Reinartz T. (1999). *Knowledge Discovery in Databases*. Springer-Verlag Heidelberg, Vol. 1623/1999.

Tsai H. T., J.W. Lane, and C.S. Lin (1986). *Modern Control Techniques for The Processing Industries*, Marcel Dekker Inc., 1986.

Wise B.M (1991). Adapting Multivariate Analysis for Monitoring and Modeling of Dynamic Systems. *PhD Dissertation*, University of Washington.

Wold S. (1978). Cross-validatory Estimation of the Number of Components in Factor and Principal Components Model. *Technometrics,* 20(4).

Zaiane O.R. (2002). Principles of Knowledge Discovery in Databases. *Course Notes*, Dpt. of Computing Science, University of Alberta.