Applying Data Mining Techniques to Address Critical Process Optimization Needs in Advanced Manufacturing

Li Zheng, Chunqiu Zeng, Lei Li,
Yexi Jiang, Wei Xue, Jingxuan Li,
Chao Shen, Wubai Zhou, Hongtai Li,
Liang Tang, Tao Li
School of Computer Science
Florida International University
{lzhen001,czeng001,taoli}@cs.fiu.edu

Bing Duan, Ming Lei,
Pengnian Wang
ChangHong COC Display Devices Co., Ltd
35 East Mianxing High-Tech Park
Mianyang, Sichuan, China 621000
{bing.duan,thunder,wpn}@changhong.com

ABSTRACT

Advanced manufacturing such as aerospace, semi-conductor, and flat display device often involves complex production processes, and generates large volume of production data. In general, the production data comes from products with different levels of quality, assembly line with complex flows and equipments, and processing craft with massive controlling parameters. The scale and complexity of data is beyond the analytic power of traditional IT infrastructures. To achieve better manufacturing performance, it is imperative to explore the underlying dependencies of the production data and exploit analytic insights to improve the production process. However, few research and industrial efforts have been reported on providing manufacturers with integrated data analytical solutions to reveal potentials and optimize the production process from data-driven perspectives.

In this paper, we design, implement and deploy an integrated solution, named `PDP-Miner`, which is a data analytics platform customized for process optimization in Plasma Display Panel (PDP) manufacturing. The system utilizes the latest advances in data mining technologies and Big Data infrastructures to create a complete analytical solution. Besides, our proposed system is capable of supporting automatically configuring and scheduling analysis tasks, and balancing heterogeneous computing resources. The system and the analytic strategies can be applied to other advanced manufacturing fields to enable complex data analysis tasks. Since 2013, `PDP-Miner` has been deployed as the data analysis platform of ChangHong COC[1]. By taking the advantages of our system, the overall PDP yield rate has increased from 91% to 94%. The monthly production is boosted by 10,000 panels, which brings more than 117 million RMB of revenue improvement per year[2].

Categories and Subject Descriptors: H.2.8[Database Applications]: Data Mining; H.4[Information Systems Applications]: Miscellaneous

Keywords: Advanced Manufacturing, Big Data, Data Mining Platform, Process Optimization

1. INTRODUCTION

The manufacturing industry involves the production of merchandise for use or sale using labor and machines, tools, chemical processing, etc. It has been the mainstay of many developed economies and remains an important driver of GDP (Gross Domestic Product). According to the Bureau of Economic Analysis data, every dollar goods in manufacturing generates $1.48 in economic activity, the highest economic multiplier among major economic sector[3]. With the advancement of new technologies, a lot of manufacturers utilize cutting-edge materials and emerging capabilities enabled by physical, biological, chemical and computer sciences. The improved manufacturing process often refers to as "advanced manufacturing" [14, 28]. For example, organizations in oil and gas industry apply new technologies to transform raw data into actionable insight to improve asset value and product yield while enhancing safety and protecting the environment.

In advanced manufacturing, a medium-sized or large manufacturing sector often arranges complex and elaborate production processes according to the product structure, and generates large volume of production data collected by sensor technologies [8], Manufacturing Execution System (MES) [13], and Enterprise Resources Planning (ERP) [22]. In practice, the production data contains intricate dependencies among a tremendous amount of controlling parameters in the production workflow. Generally, it is extremely difficult or even impossible for analysts to manually explore such dependencies, let alone proposing strategies to potentially optimize the workflow.

Fortunately, the use of data analytics offers the manufacturers great opportunities to acquire informative messages

[1]ChangHong COC Display Devices Co., Ltd is one of the world's largest display device manufacturing companies in China (http://www.cocpdp.com).

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623347.

[2]http://articles.e-works.net.cn/mes/article113579.htm.

[3]JEC Democratic staff calculations based on data from the Bureau of Economic Analysis, Industry Data, Input-Output Accounts, Industry-by-Industry Total Requirements after Redefinitions (1998 to 2011).

Table 1: Perspective Differences Between Manufacturers and Data Analysts.

	Capacity	Capability	Knowledge
Manufacturers	• huge production output • sophisticated workflow • complex supply chain	• control yield rate • optimize production line • effective parameter setting	• private Know-How • high dependency to experts • high cost of testing
Data Analysts	• large number of samples • high-dimensional data • complex param dependencies	• process optimization • feature reduction and selection • feature association analysis	• utilize domain expertise • knowledge sharing • knowledge management
Application Gap	• utilize customized data analysis algorithms to mine the underlying knowledge; • provide configurable task platforms to allow automatic taskflow execution; • enable efficient knowledge representation and management.		

towards optimizing the production workflow. However in practice, there is a significant **application gap** between manufacturers and data analysts in observing the data and using automation tools. Table 1 highlights the perspective difference between manufacturers and analysts on three important aspects: (1) *Capacity*, i.e. what the data looks like; (2) *Capability*, i.e., how the data can be utilized; and (3) *Knowledge*, i.e., how to perform knowledge discovery and management.

To bridge the gap, it is imperative to provide automated tools to the manufacturers to enhance their capability of analyzing production data. Data analytics in advanced manufacturing, especially data mining approaches, have been targeting several important fields, such as product quality analysis [20, 26], failure analysis of production [3, 25], production planning and scheduling analysis [1, 2], analytic platform implementation [7, 8], etc. However, few research and industrial efforts have been reported on providing manufacturers with an **integrated data analytical platform**, to enable automatic analysis of the production data and efficient optimization of the production process. Our goal is to provide such a solution to practically fill the gap.

1.1 A Concrete Case: PDP Manufacturing

Plasma Display Panel (PDP) manufacturing produces over 10,000 panels for a daily throughput in ChangHong COC Display Devices Co., Ltd (COC for short). The production line is near 6,000 meters and the process contains 75 assembling routines, and 279 major production equipments with more than 10,000 parameters. The average production time throughout the manufacturing process requires 76 hours. Specifically, the workflow consists of three major procedures shown in Figure 1, i.e., *front panel*, *rear panel*, and *panel assembly*. Each procedure contains multiple sequentially executed flows, and each flow is composed of multiple key routines. The first two procedures are executed in parallel, and each pair of front and rear panels will be assembled in the assembly procedure. Figure 2 depicts the real assembly line of one routine (Tin-doped Indium Oxide, ITO) in front panel procedure, which gives us a sense of how complex the complete production process will be.

There are 83 types of equipments in the PDP manufacturing process, each of which has a different set of parameters to fulfill the corresponding processing tasks. The parameters are often preset to certain values to ensure the normal operation of each equipment. However, the observed parameter values often deviate from the preset values. Further in the production environment, external factors, e.g., temperature, humidity, and atmospheric pressure, may potentially affect

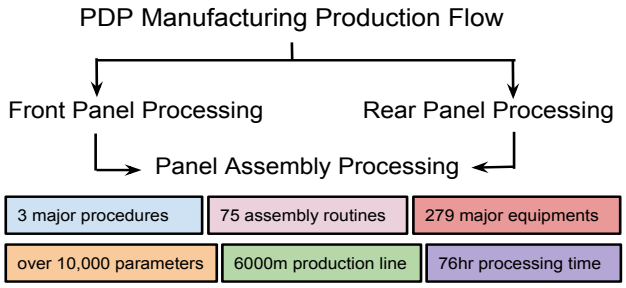


Figure 1: PDP Manufacturing Production Flow.

Figure 2: An Example Routine in PDP Workflow.

the product quality as the raw materials and equipments are sensitive to these factors. The observed values of external factors vary significantly in terms of sensor locations and acquisition time. The production process generates a huge amount of production data (10 Gigabytes per day with 30 Million records).

In daily operations, the manufactures are concerned with how to improve the yield rate of the production. To achieve this goal, several questions need to be carefully addressed, including

- What are the key parameters whose values can significantly differentiate qualified products from defective products?
- How the parameter value changes affect the production rate?
- What are the effective parameter recipes to ensure high yield rate?

Answering these questions, however, is a non-trivial task due to the scale and complexity of the production data, and is impossible for domain analysts to manually explore the data. Hence, it is necessary to automate the optimization

process using appropriate infrastructural and algorithmic solutions.

1.2 Challenges and Proposed Solutions

The massive production data poses great challenges to manufacturers in effectively optimizing the production workflow. During the past two years, we have been working closely with the technicians and engineers from COC to investigate data-driven techniques for improving the yield rate of production. During this process, we have identified two key challenges and proposed the corresponding solutions to each challenge as follows.

In general, highly automatic production process often generates large volume of data, containing a myriad of controlling parameters with the corresponding observed values. The parameters may have malformed or missing values due to inaccurate sensing or transmission. Therefore, it is crucial to efficiently store and preprocess these data, in order to handle the increasing scale as well as the incomplete status of the data. In addition, the analytics of the production data is a cognitive activity towards the production workflow, which embodies an iterative process of exploring the data, analyzing the data, and representing the insights. A practical system should provide an integrated and high-efficiency solution to support the process.

CHALLENGE 1. *Facing the enormous data with sustained growth, how to efficiently support large-scale data analysis tasks and provide prompt guidance to different routines in the workflow?*

Existing data mining products, such as Weka [9], SPSS and SQL Sever Data Tools, provide functionalities to facilitate users to conduct the analysis. However, these products are designed for small or medium scale data analysis, and hence cannot be applied to our problem setting. To address CHALLENGE 1, we design and implement an integrated *Data Analytics Platform* based on a distributed system (*FIU-Miner*) [32] to support high-performance analysis. The platform manages all the production data in a distributed environment, which is capable of configuring and executing data preprocessing and data analysis tasks in an automatic way. The platform has the following functionalities: (1) cross-language data mining algorithms integration, (2) real-time monitoring of system resource consumption, and (3) balancing the node workload in clusters.

Besides CHALLENGE 1, in advanced manufacturing, the controlling parameters in the production workflow may correlate with each other, and potentially affect the production yield rate. Several analysis tasks identified by PDP analysts include (1) discovering the most related parameters (TASK 1); (2) quantifying the parameter correlation with the product quality (TASK 2); and (3) proposing novel parameter recipes (i.e., parameter value combinations) to improve the yield rate (TASK 3). A reasonable way to effectively fulfill these tasks is to utilize suitable data mining and machine learning techniques. However, existing algorithms cannot be directly applied to these tasks, as they may either lack the capability of handling large-scale data, or fail to consider domain-specific data characteristics.

CHALLENGE 2. *Facing various types of mining requirements, how to effectively adapt existing algorithms for customized analysis tasks that comprehensively consider the domain characteristics?*

In our proposed system, CHALLENGE 2 is effectively tackled by developing appropriate data mining algorithms and adapting them to the problem of analyzing the manufacturing data. In particular, to address TASK 1, we propose an ensemble feature selection method to generate a stable parameter set based on the results of various feature selection methods. To address TASK 2, we utilize regression models to describe the relationship between product quality and various parameters. To address TASK 3, we apply association based methods to identify possible feature combinations that can significantly improve the quality of product. To make the system an integrated solution, we also provide the functionalities of data exploration (including comparative analysis and data cube) and result management.

Our proposed solution, `PDP-Miner`, is essentially a scalable, easy-to-use and customized data analysis system for large-scale and complex mining tasks on manufacturing data. Exploitation of the latest advances in data mining and machine learning technologies unleashes the potential to achieve three critical objectives, including *enhancing exploration and production*, *improving refining and manufacturing efficiency*, and *optimizing global operations*. Since 2013, `PDP-Miner` has been deployed as the production data analysis platform of COC. By using our system, the overall yield rate has increased from 91% to 94%, which has brought more than 117 million RMB of revenue per year[4].

1.3 Roadmap

The rest of the paper is organized as follows. Section 2 presents an overview of our proposed system, starting from introducing the system architecture, followed by the details of three interleaved analysis modules, including data exploration, operational analysis and result representation. In Section 3, we explore possible feature selection strategies to identify pivotal parameters in the production process, and propose an ensemble feature selection approach to obtain robust yet predominant parameter set. In Section 4, we discuss the task of measuring the importance of parameters, and utilize regression models to examine how the parameter change will affect the yield rate. Section 5 describes our strategy of mining the knowledge of data, that is, to employ discriminative analysis (e.g., association mining) to reveal the dependencies of parameters. Section 6 represents the system deployment, in which system performance evaluation is described and some important real findings are presented. Finally, Section 7 concludes the paper.

2. SYSTEM OVERVIEW

The overall architecture of `PDP-Miner` is shown in Figure 3. The system, from bottom to top, consists of two components: *Data Analytics Platform* (including *Task Management Layer* and *Physical Resource Layer*) and *Data Analysis Modules*.

Data Analytics Platform provides a fast, integrated, and user-friendly system for data mining in distributed environment, where all the data analysis tasks accomplished by *Data Analysis Modules* are configured as workflows and also automatically scheduled. Details of this module are provided in Section 2.1.

Data Analysis Modules provide data-mining solutions and methodologies to identify important production factors, in-

[4]http://articles.e-works.net.cn/mes/article113579.htm

cluding controlling parameters and their underlying correlations, in order to optimize production process. These methods are incorporated into the platform as functions and modules towards specific analysis tasks. In `PDP-Miner`, there are 3 major analytic modules: *data exploration*, *data analysis*, and *result management*. In Section 2.2, more details are provided by presenting our data mining solutions customized for PDP production data. A sample system for demonstration purpose is available at `http://bigdata-node01.cs.fiu.edu/PDP-Miner/demo.html`.

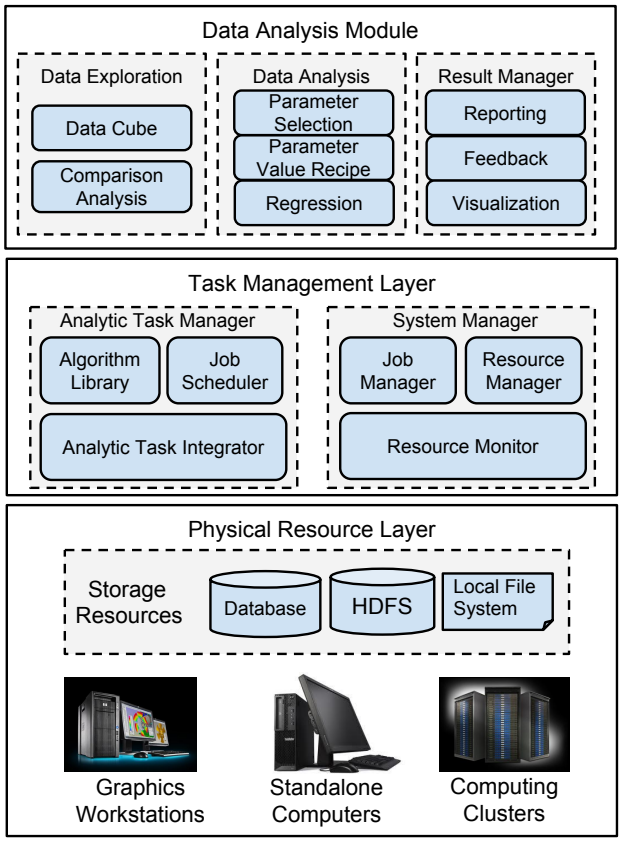


Figure 3: System Architecture.

2.1 Data Analytics Platform

Traditional data-mining tools or existing products [9, 21, 19, 18, 23, 30] have three major limitations when applied to specific industrial sectors or production process analysis: 1) They support neither large-scale data analysis nor handy algorithm plug-in; 2) They require advanced programming skills when configuring and integrating algorithms for complex data mining tasks; and 3) They do not support large scale of analysis tasks running simultaneously in heterogeneous environments.

To address the limitations of existing products, we develop the data analytic platform based on our previous large-scale data mining system, FIU-Miner [32], to facilitate the execution of data mining tasks. The data analytic platform provides a set of novel functionalities with the following significant advantages [32]:

- *Easy operation for task configuration.* Users, especially non-data-analyst, can easily configure a complex data mining task by assembling existing algorithms into a workflow. Configuration can be done through a graphic interface. Execution details including task scheduling and resource management are transparent to users.

- *Flexible supports for various programs.* The existing data mining tools, such as data preprocessing libraries, can be utilized in this platform. There is no restriction on programming languages for those programs exist or to be implemented, since our data analytic platform is capable of distributing the tasks to proper runtime environments.

- *Effective resource management.* To optimize the utilization of computing resources, tasks are executed by considering various factors such as algorithm implementation, server load balance, and the data location. Various runtime environments are supported for running data analysis tasks, including graphics workstations, stand-alone computers, and clusters.

2.2 Data Analysis Modules

2.2.1 Data Exploration

The *Comparison Analysis* and *Data Cube* are capable of assisting data analysts to explore PDP operation data efficiently and effectively.

Comparison Analysis *Comparison Analysis*, shown in Figure 6(a), provides a set of tools to help data analysts quickly identify parameters whose values are statistically different between two datasets according to several statistical indicators. *Comparison Analysis* is able to extract the top-k most significant parameters based on predefined indicators or customized ranking criteria. It also supports comparison on the same set of parameters over two different datasets to identify the top-k most representative parameters of two specified datasets.

Data Cube *Data Cube*, shown in Figure 6(b), provides a convenient approach to explore high dimensional data so that data analysts can have a glance at the characteristics of the dataset. In addition, *Data Cube* can conduct multi-level inspection of the data by applying OLAP techniques. Data analysts can customize a multi-dimensional cube over the original data. Thus, the constructed data cubes allow users to explore multiple dimensional data at different granularities and evaluate the data using pre-defined measurements.

2.2.2 Data Analysis

The data mining approaches in algorithm library can be organized as a configurable procedure in *Operation Panel*, as shown in Figure 6(c). The *Operation Panel* is a unified interface to build a workflow for executing such task automatically. The *Operation Panel* contains the following three main tasks:

Important Parameter Selection By modeling the important parameter discovery task as a feature selection problem, several feature selection algorithms are implemented adaptively based on the production data. Moreover, an advanced ensemble framework is designed to combine multiple feature selection outputs. Based on these implementations, the system is able to generate a list of important parameters, shown in Figure 6(d).

Regression Analysis The purpose of *Regression Analysis* (shown in Figure 6(f)) is to discover the correlations between the yield rate and the controlling parameters. The regression model not only indicates whether a correlation exists between a parameter and the yield rate but also quantifies the extent that the change of the parameter value will influence the yield rate.

Discriminative Analysis Discriminative analysis (See Figure 6(e)) is an alternative approach to identify the feature values that have strong indication to the target labels (panel grade). By grouping and leveraging the features of individual panels, this approach is able to find the most discriminative rules (a set of features with the values) to the target labels according to the data.

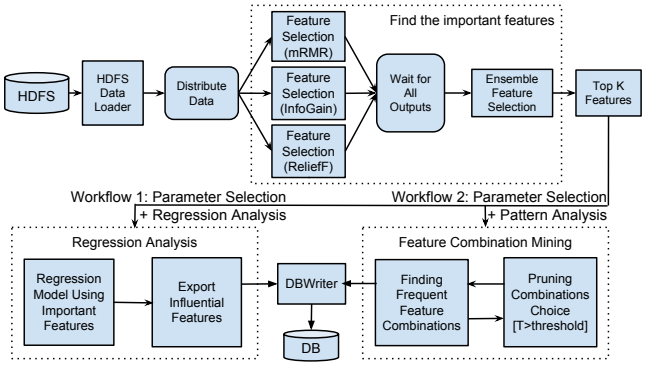


Figure 4: A Sample Workflow for PDP Manufacturing Data Analysis.

To illustrate how *Data Analysis Modules* are incorporated with the *Data Analytics Platform*, Figure 4 illustrates two example analytic tasks wrapped as two workflows. As shown, *Workflow 1* indicates an analysis procedure of building regression models with selected important parameters; *Workflow 2* indicates another procedure of identifying reasonable parameter value combinations based on previously selected parameters. The *Operation Panel* provides a user-friendly interface shown in Figure 5 to facilitate workflow assembly and configuration. Users only need to explicitly create tasks dependencies before the workflow executing automatically by our platform.

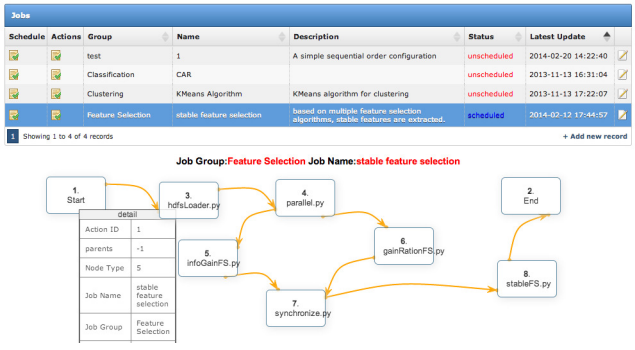


Figure 5: Data Analysis Workflow Configuration.

2.2.3 *Result Management*

The analytic results are being categorized into three types: the important parameter list, the parameter value combinations, and the regression model. Templates are designed to support automatically storage, update, and retrieval of discovered patterns. Results are recorded based on analysis tasks and can be organized in terms of important equipment, top parameters, and task list. For each result, corresponding domain experts can refine and give feedback, shown in Figure 6(h). In addition, visualizations are provided to summarize the analytic results, collected feedbacks, and status of current knowledge (shown in Figure 6(g)). It provides a flexible interface for maintaining domain knowledge very efficiently.

3. ENSEMBLE FEATURE SELECTION

In manufacturing management, the primary goal is to improve the yield rate of products by optimizing the manufacturing workflow. To this end, one important question is to identify the key parameters (features) in the workflow, which can significantly differentiate qualified products from defective ones. However, it is a non-trivial task to select a subset of features from the huge feature space. To tackle this problem, we initially experimented several widely used feature selection approaches. Specifically, we use Information Gain [10], mRMR [5] and ReliefF [24] to perform parameter selection. Figure 7 shows the top 10 selected features by these three algorithms on a sampled PDP dataset.

As observed in Figure 7, the three feature subsets share only one common feature ("Char_020101-008"). Such a phenomenon indicates the instability of feature selection methods, as it is difficult to identify the importance of a feature from a mixed view of feature subsets. In general, the selected are the most relevant to the labels and less redundant to each other based on certain criteria. However, the correlated features may be ignored if we select a small subset of features. In terms of knowledge discovery, the selected feature subset is insufficient to represent important knowledge about redundant features. Further, different algorithms select features based on different criteria, which renders the feature selection result instable.

Information gain(top10)	mRMR(top10)	ReliefF(top10)
Char_110101-004	Char_020101-016	Char_110101-009
Char_110101-003	Char_020101-004	Char_110101-008
Char_110101-005	Char_020101-008	Char_100102-079
Char_110101-002	Char_020101-009	Char_100101-199
Char_110101-006	Char_020101-010	Char_100101-208
Char_110101-001	Char_020101-007	Char_100101-212
Char_020101-008	Char_020101-006	Char_100102-013
Char_020101-017	Char_020101-003	Char_100101-213
Char_100101-168	Char_020101-014	Char_100102-081
Char_020101-013	Char_020101-013	Char_020101-008

Figure 7: Selected Features by Different Algorithms.

The stability issue of feature selection has been studied recently [4, 12] under the assumption of small sample size. The results of these work indicate that different algorithms with equal classification performance may have a wide variance in terms of stability. Another direction of stable feature selection involves exploring the consensus information among different feature groups [17, 29, 31], which first identifies consensus feature groups for a given dataset, and then performs selection from the feature groups. However, these methods fail to consider the correlation between selected features and unselected ones, which might be important to guide us for feature selection.

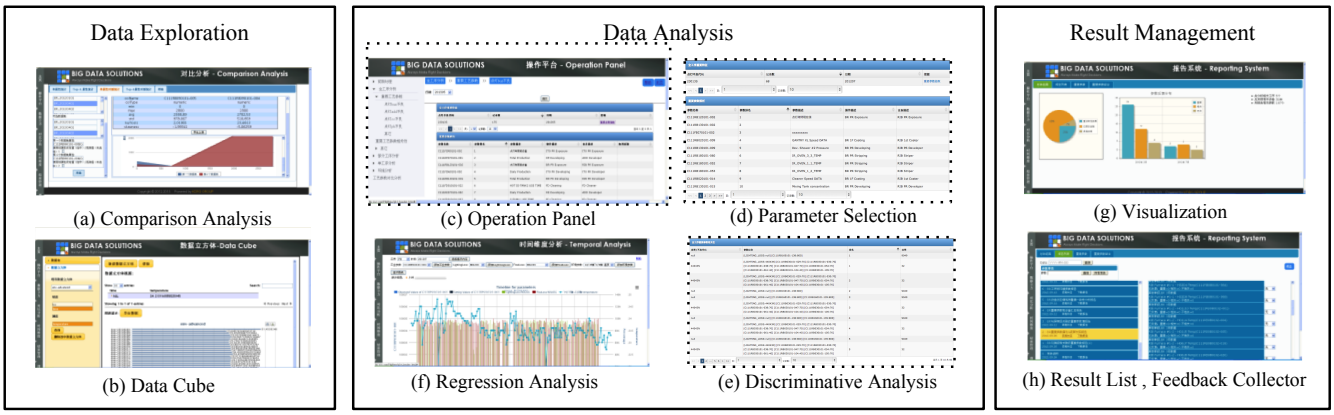


Figure 6: PDP-Miner Analysis Modules.

In our system, inspired by ensemble clustering [27, 15], we employ the ensemble strategy on the results of various feature selection methods to maintain the robustness and stability of feature selection. The problem setting of stable feature selection is defined as follows. Given a dataset with M features, we employ N feature selection methods which for an arbitrary feature i return a N-lengthed vector $\mathbf{y}_i, i = 1, 2, \ldots, M$. Each entry of $\mathbf{y}_i$ is 1 or 0 indicating whether the feature i is selected or not by the corresponding feature selection method. Since we are concerned with whether to select a feature or not, we assume a feature i, in the form of results of N feature selection methods, $\mathbf{y}_i$, is generated from a mixture of two multivariate components, indicating selected features and unselected features, i.e.,

$$p(\mathbf{y}_i) = \sum_{j=1}^{2} \pi_j p(\mathbf{y}_i|\boldsymbol{\theta}_j), \tag{1}$$

where π_j denotes the mixture probability of j-th component parameterized by $\boldsymbol{\theta}_j$, in which the n-th entry θ_{jn} means the probability of the output of n-th feature selection method equals to 1. We further assume conditional independence between feature selection methods. Therefore,

$$p(\mathbf{y}_i|\boldsymbol{\theta}_j) = \prod_{n=1}^{N} p(y_{in}|\theta_{jn}). \tag{2}$$

As the result of a feature selection method in the vector $\mathbf{y}_i$ is either selecting (1) or not selecting (0) the feature i, the probability of the feature i being selected by the n-th feature selection method, i.e., $p(y_{in}|\theta_{jn})$, could be represented by a Bernoulli distribution

$$p(y_{in}|\theta_{jn}) = \theta_{jn}^{y_{in}}(1-\theta_{jn})^{1-y_{in}}. \tag{3}$$

In addition, we assume that all the features are *i.i.d.* Then the log-likelihood of the unified probabilistic model is

$$\mathcal{L} = \sum_{i=1}^{M} \log \sum_{j=1}^{2} \pi_j p(\mathbf{y}_i|\boldsymbol{\theta}_j). \tag{4}$$

To learn the parameters π_j and $\boldsymbol{\theta}_j, j \in \{1, 2\}$, we use Expectation-Maximization (EM) algorithm. To this end, we introduce a series of hidden random variables $z_i, i = 1, 2$ to indicate $\mathbf{y}_i$ belonging to each component, i.e., the parameters of the random variable $z_{i1}, z_{z2}, z_{i1} + z_{i2} = 1$.

The iterative procedure of EM will be terminated when the likelihood of the mixture model does not change too much constrained by a predefined threshold. The hidden variable z_i indicates the probabilities of membership of feature $\mathbf{y}_i$ with respect to all mixture components. It is in some sense similar to the situation in Gaussian mixture models. The feature is assigned to the j-th component that the corresponding value z_{ij} is the largest in $z_{ij}, j \in \{1, 2\}$. As a feature selection method will eventually generate two subsets of features (selected or not), it is reasonable to make two mixture components.

After obtaining the assignments of features to components, say $\phi(z_i)$, we group features into two categories, i.e., selected/unselected. In practice, the number of selected features are significantly less than the unselected ones, and hence the features that are not selected by any feature selection method are put into a large category. The features in the other category are final feature selection results. Specifically for each component j, we pick the features that have the membership assignment, i.e., z_{ij}, greater than a predefined threshold τ, and then put these features into the selected category. In this way, we can discard features with low probabilities for selection, and hence the stability of feature selection can be achieved by assembling different feature selection results using the mixture model.

4. REGRESSION ANALYSIS

To optimize the production process, it is imperative to discover the parameters that have significant influence on the yield rate and quantify such influence. In our system, an actionable solution is to explicitly establish a relationship between controlling parameters and the yield rate, which can be achieved using regression analysis.

Formally, assume the daily observations are *i.i.d.* Then the relationship between features (parameters) and the yield rate can be modeled as a function $f(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise, i.e.,

$$y = f(\boldsymbol{x}, \boldsymbol{w}) + \epsilon, \epsilon \sim \mathcal{N}(0, \beta^{-1}), \tag{5}$$

where y denotes the yield rate, $\boldsymbol{x} = (x_1, \cdots, x_d)^T$ denotes the set of features that may have impact on y, and $\boldsymbol{w}$ denotes the weight of features. The noise term is a zero mean Gaussian random variable with precision β.

In our system, we implement two linear regression based models: *ridge regression* and *lasso regression* [11]. From the perspective of maximum likelihood, the linear relationship

can be expressed as

$$\ln p(y|\boldsymbol{w},\beta) = \sum \ln \mathcal{N}(y_i|\boldsymbol{w}^T\boldsymbol{x}_i, \beta^{-1}). \quad (6)$$

For both models, we levérage least square to quantify the error, i.e.,

$$E(\boldsymbol{w}) = \begin{cases} \frac{1}{2}\sum_i (y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \frac{1}{2}\lambda||\boldsymbol{w}||_2^2, \ \textit{ridge regression} \\ \frac{1}{2}\sum_i (y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \frac{1}{2}\lambda||\boldsymbol{w}||_1, \ \textit{lasso regression} \end{cases}. \quad (7)$$

In advanced manufacturing domain, the number of features is usually large (in PDP scenario, the number of features is more than 10K), and therefore ensemble feature selection (described in Section 3) is applied before building the regression model. To conduct the regression, we incorporate three categories of features:

1. The parameters of the equipments involved in the manufacturing process. This category of features is collected from the log of the equipments.

2. The parameters of the environment, such as temperature, humidity, and pressure, etc. This category of the features is collected from the deployed sensors in each workshop.

3. The features of the materials, such as the viscosity, consistency, and concentration, etc. This category of feature is collected from material descriptions and reports.

After integrating all the features, we normalize each dimension of the features using standardization, i.e. $\frac{x-\bar{X}}{std(X)}$.

The linear regression can be solved efficiently. When the dataset is small, the closed form can be directly obtained, i.e. $\hat{\boldsymbol{w}}^{ridge} = (\lambda\boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ for ridge regression and $\hat{\boldsymbol{w}}^{lasso} = sgn((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)(|(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}| - \lambda)$ for lasso regression, where $\boldsymbol{X}$ denotes the matrix of the features with ith row indicating the feature set $\boldsymbol{x}_i$. For large datasets, we train the model iteratively by using stochastic gradient descent for ridge regression and coordinate gradient for lasso regression.

The weights of the trained model can be intuitively interpreted. Firstly, the value $|\boldsymbol{w}_i|$ indicates the conditional correlation between the feature x_i and the yield rate given the other features. In general, a larger weight indicates a larger conditional correlation. Moreover, the corresponding p-value of each feature can be leveraged to measure the likelihood of the correlation. The smaller the p-value, the less likely such correlation is false.

By performing regression analysis on the PDP data, we find some interesting correlations. For example:

1. The variance of the humidity of the air has positive correlation wtih the yield rate. This provides empirical evidence to support the conjecture of PDP technicians that the variance of the humidity plays an important role in affecting the yield rate.

2. The pressure of the air has positive correlation with the yield rate, whereas its variance changes inversely. The less the pressure changes, the higher the yield rate would be.

3. The workshop temperature and its variance vary slightly within a small range, and the corresponding weight is very small. In practice, the change of the temperature may affect the usage of materials as well as the production process. Hence, it is often being carefully controlled by technicians.

5. DISCRIMINATIVE ANALYSIS

Discriminative analysis mines the feature knowledge of the PDP panel data from a different perspective. It is used as an alternative way to reveal the underlying relationship between the features and the panel grades. Specifically, it helps discover parameter recipes as well as sets of feature values which are closely related to qualified panels and defective panels. In `PDP-Miner`, the techniques of *association based classification* [16] and low-support discriminative pattern mining [6] are leveraged to conduct the discriminative analysis.

5.1 Association based classification

Association based classification integrates classification and association rule mining to discover a special subset of association rules called *class association rules (CARs)*. A CAR is an implication of the form $\{r : F \rightarrow y\}$, where F is a subset of the entire feature value set and y denotes the class label (the PDP panel grade in our scenario). Each CAR is associated with support s and confidence c, indicating how many records contain F and the ratio of records containing F that are labeled as y. In general, CARs contain strong discriminative information to infer the PDP panel grades. A rule-based classifier can be built by selecting a subset of the CARs that collectively cause the least error, i. e. $r_1, r_2, ..., r_n \rightarrow y$.

Compared with feature selection and regression analysis, association based classification enables the possibility of early detection due to the unique characteristics of CARs. If CARs only refer to the features in the early manufacturing process, this method can quickly identify semi-finished yet defective panels, and prevent further resource waste.

The early detection strategy is useful in the advanced manufacturing domain, as any earlier detected bad semi-finished product can directly reduce the manufacturing cost. For the production with a large number of assembling procedures, such a reduction is not trivial.

5.2 Low support discriminative pattern mining

A manufacturing process could consist of hundreds of assembling procedures with thousands of tunning parameters. When the feature dimension is high, standard association rule based methods would become time-consuming. A naïve solution for this scenario is to increase the support threshold to speed up mining. However, this strategy may miss interesting low-support patterns.

To address this problem, we adapt the idea of low support pattern mining algorithm (*SMP*) [6] and integrate the algorithm into `PDP-miner`. *SMP* aims at mining the discriminative patterns by leveraging a family of anti-monotonic measures called *SupMaxK*. *SupMaxK* organizes the discriminative pattern set into nested layers of subsets.

5.2.1 Discriminative Patterns Detection

Many association mining methods utilize "support" to select rules/patterns. Different from the traditional association mining, the "discriminative support" is defined to mea-

sure the quality (discriminative capability) of the rule set:

$$DisS(\alpha) = |S_{qualified}(\alpha) - S_{defective}(\alpha)|, \quad (8)$$

where α is a set of parameter values, $S_{qualified}$ and $S_{defective}$ denote the "support" of α over two classes, indicating whether the target panel is qualified or defective.

A naïve implementation using this measure suffers from low efficiency [6] when pruning frequent non-discriminative rules. To address this issue, a new measure – $SupMaxK(\alpha)$ is introduced to help prune unrelated patterns by estimating $S_{defective}(\beta)$.

$$SupMaxK(\alpha) = S_{qualified}(\alpha) - max_{\beta \in \alpha}(S_{defective}(\beta)), \quad (9)$$

where $|\beta| = K$, β is the subset of α. Three reasons make this measure useful: (1) $SupMaxK$ can help select more discriminative patterns as K increases; (2) $SupMaxK$ is a lower bound of $DisS$; (3) $SupMaxK$ is anti-monotonic.

Due to the anti-monotonic property of $SupMaxK$, *SMP* can naturally be utilized to mine the discriminative patterns whose support are low but have strong indication to the panel grades.

6. SYSTEM DEPLOYMENT

We evaluate our proposed system from two aspects: the system performance and the real findings. The evaluation demonstrates that our system is a practical solution for large-scale data analysis, through integrating and adapting classic data mining techniques and customizing them for specific domains, particularly, advanced manufacturing.

6.1 System performance

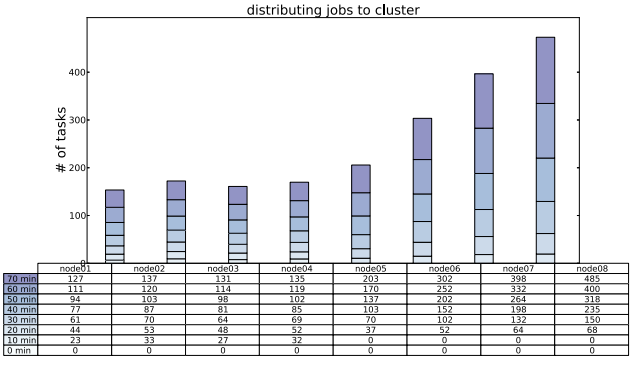
Our system is able to perform large-scale data analysis and can be easily scale up. To demonstrate the scalability of `PDP-Miner`, we design a series of cluster workload balance experiments in both static and dynamic computing environments. The experiments are conducted on a testbed cluster separated from the real production system. The cluster consists of 8 computing nodes with different computing performances.

In the experiments, one frequent analysis task of `PDP-Miner` is created using the job configuration interface, which consists of two sequential functions, i.e., Parameter Selection → Parameter Combination Extraction. For evaluation purpose, ten different datasets (about 30 million records) are generated by sampling from the original 1-year production datasets. The analysis task is conducted over these datasets in two types of experiments: *Exp I Workload balance in a static environment* and *Exp II Workload balance in a dynamic environment*. In the following, we describe the detailed experimental plans as well as the results.

Exp I: Each node in the cluster is deployed with one *Worker*. We configure 10 parameter selection tasks with different running times in `PDP-Miner`. Each job starts at time 0 and repeats with a random interval (< 1 minutes). Figure 8 shows how our system balances the workloads based on the underlying infrastructures. The x-axis denotes the time and the y-axis denotes the average number of completed jobs for each *Worker* at the given moment during the task execution. Clearly, the accumulated number of completed jobs (the blue solid bars) increases linearly, whereas the amortized number of completed jobs (the white empty bars) remains stable. This shows that when the cluster remains unchanged, our system achieves a good balance of the resource utilization by properly distributing jobs. The effective distribution of jobs guarantees a full use of existing resources to maximize the throughput without incurring resource bottleneck.

Figure 8: Load Balance in Static Environment.

Exp II: To investigate the resource utilization of `PDP-Miner` under a dynamic environment, we initially provide four nodes (node1∼4), each with 1 *Worker*, and then add the other four nodes (node5∼8) 10 minutes later. To emulate the nodes with different computing powers, the newly-added nodes are deployed with 2 to 5 *Workers*, respectively. Each *Worker* is restricted to use only 1 CPU core at a time, so the node deployed with more *Workers* can have more powerful computing resources. Figure 9 shows the number of jobs completed by each node during observing the system execution for 70 minutes. The number of jobs on each node is segmented every 10 minutes. It clearly shows that the number of completed jobs is proportional to the number of *Workers* on each node, which indicates that our system can balance the workloads in a dynamically changed cluster. It also demonstrates that the entire system can be linearly extended with resources of different computing power.



	node01	node02	node03	node04	node05	node06	node07	node08
70 min	127	137	131	135	203	302	398	485
60 min	111	120	114	119	170	252	332	400
50 min	94	103	98	102	137	202	264	318
40 min	77	87	81	85	103	152	198	235
30 min	61	70	64	69	70	102	132	150
20 min	44	53	48	52	37	52	64	68
10 min	23	33	27	32	0	0	0	0
0 min	0	0	0	0	0	0	0	0

Figure 9: Load Balance in Dynamic Environment.

6.2 Real Findings

`PDP-Miner` has been playing an important role on revealing deeper and finer relations behind big data in COC's real practice. As an example, WorkFlow1 in Figure 4 is executed to extract important parameters from a single procedure, named barrier-rib (BR). 30 selected parameters are reported

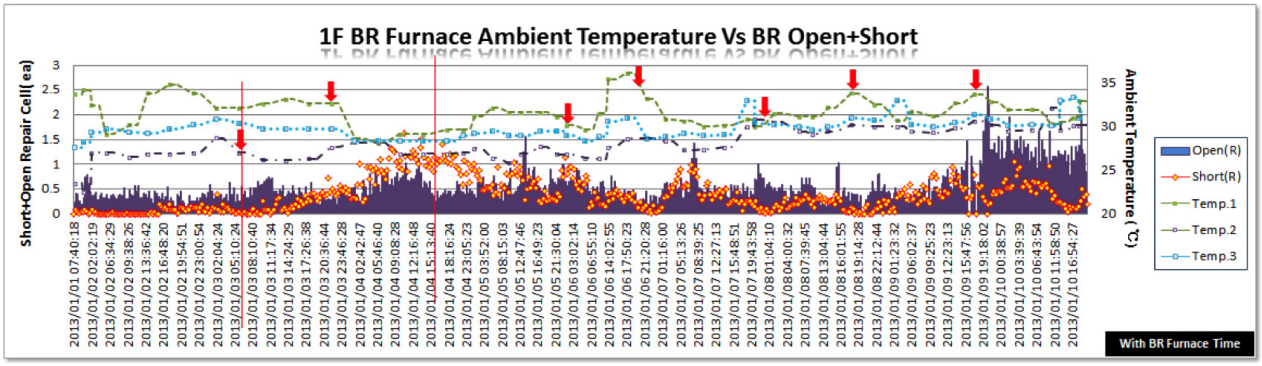


Figure 10: Real Case of Regression Analysis Results.

and verified by domain experts. Within these 30 parameters, 15 of them have already been carefully monitored by the analysts, which is consistent with domain knowledge. Another 9 parameters, which are not monitored in the previous production, are confirmed to have great impact on the product quality. After applying WorkFlow1 to the entire production data, 197 important parameters are reported by our system, among which 133 parameters are consistent with production experience, and 50 parameters are verified by domain experts to have direct impact on the product quality. The details are shown in Figure 11 (blue portion $\sim$ consistent with domain expertise; red portion $\sim$ confirmed to be important which was previously ignored; white portion $\sim$ excluded after verification).

To discover meaningful parameter values, WorkFlow2 in Figure 4 is used. We separate the production data to two sets by the product qualities (GOOD, i.e., qualified products, and SCRAP, i.e., defective products) and execute WorkFlow2 on these two sets, respectively. The analysis generates hundreds of frequent parameter value combinations for each given dataset (the number of outputs can be restricted by empirically setting a threshold of confidence). By extracting the frequent combinations in SCRAP that are not frequent in GOOD, we can obtain the value combinations that may result in defective products. Figure 12 shows a verification of a sample combination <para-xxxx-014=0, para-xxxx-015=0 or 24, para-xxxx-043=44 or 48> (big red crosses indicate that the values present densely on SCRAP products). Such a parameter value combination should be avoided in the production practice.

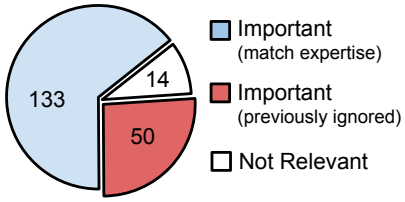


Figure 11: Important Parameters Discovered.

By applying regression analysis in WorkFlow1 of Figure 4, we discovered that environmental parameters, such as temperature and humidity, have significant correlations with the product quality. Further analysis confirmed that when the surrounding temperature of BR Furnace is under 27 °C, the

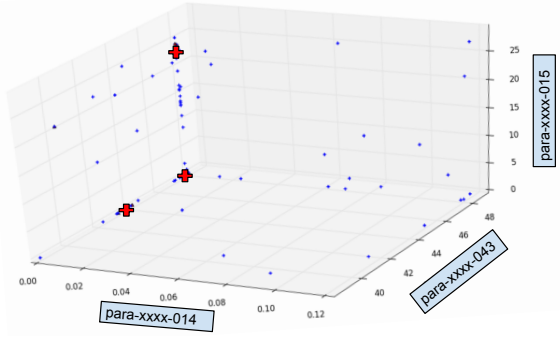


Figure 12: A Sample Parameter Combination.

number of defective products with BR Open or BR Short increases dramatically. Figure 10 depicts such findings.

The aforementioned findings are some typical examples obtained from the practical usage of our proposed system. Most of our findings have been validated by PDP technicians and are incorporated into their operational manual.

6.3 Deployment Practice

`PDP-Miner` has been successfully applied in ChangHong COC's PDP production line of the 3rd and 4th generations of products for manufacturing optimization. Every time the product line is upgraded, the yield rate drops significantly since previous parameter settings could not match new products requirements. The earlier parameters are tuned properly, the greater the cost will be reduced. `PDP-Miner` has been intensively used in such situations for problem diagnosis, including quickly identifying problematic parameter settings, detecting abnormal parameter values, and monitoring sensitive parameters.

In summary, our system brings several great benefits in optimizing the production process:

- Through establishing the relationship between parameter settings and product quality, manufacturers are more confident to properly control the production process based on analytical evidence. The cost has been greatly reduced as the number of defective products decreases.

- The prompt analysis of the production data enables the quick diagnosis on parameter values, especially when upgrading the assembly line or handling unexpected faults. As a result, the throughput increases.

- A knowledge database is constructed to manage useful analytic results that have been verified and validated by existing domain expertise. Technicians can refer to the database to look for possible solutions and control the assembly line more efficiently.

By taking advantage of our system, the overall PDP yield rate increases from 91% to 94%. Monthly production capacity is boosted by 10,000 panels, which brings more than 117 million RMB of revenue improvement per year[5]. Our system plays an revolutionary role and can be naturally transferred to other flat panel industries, such as Liquid Crystal Display (LCD) panels and Organic Light-Emitting Diode (OLED) panels, to generate great social and economic benefits.

7. CONCLUSION

`PDP-Miner` has been deployed as an important supplementary component since the year 2013. It enables prompt data analysis and efficient knowledge discovering in advanced manufacturing processes. The improved production efficacy shows that a practical data-driven solution that considers both system flexibility and algorithm customization is expected to fill the application gap between the manufacturer and data analysts. We firmly believe that, if properly being applied, the use of data analytics will become a dominating factor to underpin new waves of productivity growth and innovation, and to transform the way of manufacturings across industries in a fundamental manner.

Acknowledgement

This project is supported partially by National Science Foundation (NSF) under Grants CNS-1126619 and IIS-1213026, U.S. Department of Homeland Security under Awards 2010-ST-062-000039 and 2009-ST-061-CI0001 (VACCINE Center), Army Research Office under grants W911NF-10-1-0366 and W911NF-12-1-0431, and an FIU Dissertation Year Fellowship.

8. REFERENCES

[1] R. Belz and P. Mertens. Combining knowledge-based systems and simulation to solve rescheduling problems. *Decision Support Systems*, 17(2):141–157, 1996.

[2] I. Chen. Planning for erp systems: analysis and future trend. *Business process management journal*, 7(5):374–386, 2001.

[3] W. Chen, S. Tseng, and C. Wang. A novel manufacturing defect detection method using association rule mining techniques. *Expert systems with applications*, 29(4):807–815, 2005.

[4] C. Davis, F. Gerick, V. Hintermair, C. Friedel, K. Fundel, R. Küffner, and R. Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006.

[5] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[6] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *TKDE*, 24(2):279–294, 2012.

[7] C. Groger, F. Niedermann, H. Schwarz, and B. Mitschang. Supporting manufacturing design by analytics, continuous collaborative process improvement enabled by the advanced manufacturing analytics platform. In *CSCWD*, pages 793–799. IEEE, 2012.

[8] C. Gröger, F. Niedermann, and B. Mitschang. Data mining-driven manufacturing process optimization. In *Proceedings of the World Congress on Engineering*, volume 3, pages 4–6, 2012.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 2009.

[10] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2011.

[11] T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning*, volume 2. 2009.

[12] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[13] J. Kletti. *Manufacturing Execution Systems (MES)*. Springer, 2007.

[14] D. Lei, M. A Hitt, and J. D Goldhar. Advanced manufacturing technology: organizational design and strategic flexibility. *Organization Studies*, 17(3):501–523, 1996.

[15] T. Li and C. Ding. Weighted ensemble clustering. In *SIAM Data Mining*, 2008.

[16] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *SIGKDD*, 1998.

[17] S. Loscalzo, L. Yu, and C. Ding. Consensus group stable feature selection. In *SIGKDD*, pages 567–576, 2009.

[18] MILK. http://pythonhosted.org/milk.

[19] MLC++. http://www.sgi.com/tech/mlc.

[20] S. Oh, J. Han, and H. Cho. Intelligent process control system for quality improvement by data mining in the process industry. In *Data mining for design and manufacturing*, pages 289–309. Springer, 2001.

[21] S. Owen, R. Anil, T. Dunning, and E. Friedman. *Mahout in Action*. Manning, 2011.

[22] R. Palaniswamy and T. Frank. Enhancing manufacturing performance with erp systems. *Information systems management*, 17(3):43–55, 2000.

[23] Z. Prekopcsak, G. Makrai, T. Henk, and C. Gaspar-Papanek. Radoop: Analyzing big data with rapidminer and hadoop. In *RCOMM*, 2011.

[24] M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69, 2003.

[25] L. Shen, F. EH Tay, L. Qu, and Y. Shen. Fault diagnosis using rough sets theory. *Computers in Industry*, 43(1):61–72, 2000.

[26] V. A Skormin, V. I Gorodetski, and L. J Popyack. Data mining technology for failure prognostic of avionics. *TAES*, 38(2):388–403, 2002.

[27] A. Topchy, A. K Jain, and W. Punch. A mixture model of clustering ensembles. In *SDM*, pages 379–390, 2004.

[28] T. D Wall, J. M Corbett, R. Martin, C. W Clegg, and P. R Jackson. Advanced manufacturing technology, work design, and performance: A change study. *Journal of Applied Psychology*, 75(6):691, 1990.

[29] A. Woznica, P. Nguyen, and A. Kalousis. Model mining for robust feature selection. In *SIGKDD*, pages 913–921, 2012.

[30] L. Yu, J. Zheng, B. Wu, B. Wang, C. Shen, L. Qian, and R. Zhang. Bc-pdm: Data mining, social network analysis and text mining system based on cloud computing. In *SIGKDD*, 2012.

[31] L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *SIGKDD*, pages 803–811, 2008.

[32] C. Zeng, Y. Jiang, L. Zheng, J. Li, L. Li, H. Li, C. Shen, W. Zhou, T. Li, B. Duan, M. Lei, and P. Wang. FIU-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment. In *SIGKDD*, 2013.

[5]http://articles.e-works.net.cn/mes/article113579.htm.