Chin. J. Mech. Eng. (2017) 30:1216–1226
DOI 10.1007/s10033-017-0179-0

ORIGINAL ARTICLE

MapReduce Based Parallel Bayesian Network for Manufacturing Quality Control

Mao-Kuan Zheng[1] · Xin-Guo Ming[1] · Xian-Yu Zhang[1] · Guo-Ming Li[1]

Received: 4 December 2016 / Revised: 26 April 2017 / Accepted: 23 July 2017 / Published online: 3 August 2017

Abstract Increasing complexity of industrial products and manufacturing processes have challenged conventional statistics based quality management approaches in the circumstances of dynamic production. A Bayesian network and big data analytics integrated approach for manufacturing process quality analysis and control is proposed. Based on Hadoop distributed architecture and MapReduce parallel computing model, big volume and variety quality related data generated during the manufacturing process could be dealt with. Artificial intelligent algorithms, including Bayesian network learning, classification and reasoning, are embedded into the Reduce process. Relying on the ability of the Bayesian network in dealing with dynamic and uncertain problem and the parallel computing power of MapReduce, Bayesian network of impact factors on quality are built based on prior probability distribution and modified with posterior probability distribution. A case study on hull segment manufacturing precision management for ship and offshore platform building shows that computing speed accelerates almost directly proportionally to the increase of computing nodes. It is also proved that the proposed model is feasible for locating and reasoning of root causes, forecasting of manufacturing outcome, and intelligent decision for precision problem solving. The integration of bigdata analytics and BN method offers a whole new perspective in manufacturing quality control.

Supported by 2015 Special Funds for Intelligent Manufacturing of China MIIT (Grant No. 2015-415), and National Natural Science Foundation of China (Grant No. 71632008).

✉ Xin-Guo Ming
xgming@sjtu.edu.cn

Mao-Kuan Zheng
zhengmaokuan@163.com

[1] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Keywords Bayesian network · Big data analytics · MapReduce · Quality control

1 Introduction

In the age of Industrial 4.0, manufacturing industry is developing to digital, networked and intelligent models. Process management and control rely more on intelligent decision-making approaches [1]. With technology improvement in Internet of Things (IoT) and sensor network, massive dynamic data including environmental parameters, positions, device status, quality data, etc., can be obtained in real-time. However, as the volume of data increases exponentially without effective dealing approaches, the huge storage has become a heavy burden for those manufacturing enterprises. Solving the storage problem and mining potential values have become a hot topic in both academia and industry areas.

Quality control is the key component of manufacturing industry. Approaches and methods for quality control have been developing along with the improvement of techniques. Quality control is a systematic problem, which involves multiple factors of environment, equipment, humans, materials, designing, processes, and so on. Quality problems have more dynamic and uncertainty with the disturbance of multiple factors. With the increasing complexity of products and processes, manufacturing systems and supply chains, conventional quality management approaches are no longer able to meet production requirements. Statistical analysis methods [2] based on sampling, such as Statistical Process Control (SPC), has been widely used for quality control, because of its practical operability and simplicity. However, the process of

sampling would have information distortion, as only parts of data are used to represent of the entity.

Big data analytics can be helpful to support global manufacturing and supply chain innovation by creating data transparency, improving human decision-making and promoting innovative business models [3]. However, it still lacks available and effective data analytics techniques to help decision makers and managers to capture and harvest the potential value of data [4]. Big data has attracted great attention and been widely applied to network consumption pattern analysis, infectious diseases control and many other areas. But applications in manufacturing industry, especially in quality control, are still in the initial stage [5, 6]. The latest big data technologies are supposed to provide new effective ways for manufacturing quality control in exploration.

As mentioned before, big data acquired from manufacturing process are huge treasures. But some obstacles still exit, which hinder us from further usage and deep mining for data value. The problems mainly include:

(1) Data acquisition of quality data in many process remains manual operation. The integrity, accuracy and reliability need to be improved.
(2) The variety of data acquisition sources leads to the variety of data types. The mixture of structured, semi-structured and unstructured data is hard to get handled with traditional statistics approaches.
(3) Lack of methods and approaches for big data analytics and application in manufacturing industry. Data are simply stored rather than used.

For problems of (2) and (3), the proposal of distributed storage systems and parallel computing framework for big data provides a possible solution. Bigtable [7], Google File System (GFS) [8], and the Hadoop Distributed File System (HDFS) [9] are popular used distributed storage systems. Meanwhile, MapReduce was proposed as a programming model, for the parallel computing of large data sets (more than 1 TB) [10]. Solutions for problem (1) also can be found in existing works. Theories of grey system [11], rough sets [12], complex network [13] and Bayesian networks (BN) [14] were proposed for reasoning under uncertainties. Because of the visualization capability, system knowledge mining ability and probability characteristic, BN has become the preferred method to solve problem (1). Based on existing works, the authors try to offer a possible comprehensive solution for manufacturing quality control integrating BN and big data analytics.

The rest of the contents will be organized as follows: Related works about BN and MapReduce are clarified in Section 2. Then a framework integrating big data analytics and BN learning is proposed in Section 3. BN methods and processes for manufacturing quality data control are investigated in Section 4. After that, a comprehensive BN learning approaches integrating K2 score, EM and ERV based on MapReduce platform are studied in Section 4. A case study of hull segment manufacturing precision control in shipbuilding industry is discussed for testing and validation in Section 5. Finally come to the conclusions in the last section.

2 Related Work

2.1 Bayesian Networks

A Bayesian Network (BN) is a directed acyclic graph in which every node represents a random variable with a discrete or continuous state [15, 16]. The relationships among variables, pointed out by arcs, are interpreted in terms of conditional probabilities according to Bayes theorem [17].

With the BN is implemented the concept of conditional independence that allows the factorization of the joint probability, through the Markov property, in a series of local terms that describe the relationships among variables:

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i | pa(x_i)), \tag{1}$$

where $pa(x_i)$ denotes the states of the predecessors (parents) of the variable (child) [15, 16, 18].

Based on BN, Dynamic Bayesian Network (DBN) was developed taking the time factor into account. A DBN is still a Bayesian Network, but further relates variables to each other over adjacent time steps. DBN is often called a Two-Time slice BN because at any point in time T, network parameters can be calculated from the internal regressors and the immediate prior value (time $T-1$). DBNs have shown potential for a wide range of data mining applications by integrating priori and posterior knowledge [16].

The initial knowledge is represented in the form of a prior probability distribution over model structures and parameters, and updated using the data to obtain a posterior probability distribution over models and parameters. More formally, assuming a prior distribution over models structure $P(M)$ and a prior distribution over parameters for each model structure $P(\theta|M)$, a data set D is used to form a posterior distribution over models using Bayes rules

$$P(M|D) = \frac{\int P(D|\theta, M) P(\theta|M) \mathrm{d}\theta P(M)}{P(D)}, \tag{2}$$

which integrates the uncertainty in the parameters. For a given model structure, we can compute the posterior distribution over the parameters:

$$P(\theta|M,D) = \frac{P(D|\theta,M)P(\theta|M)}{P(D|M)}. \quad (3)$$

When comes to the quality factor modelling of manufacturing process, the data set is some sequence observations $D = \{Y_1, Y_2, \ldots, Y_T\}$, and we wish to predict the next observation, Y_{T+1} based on the data and models, then the Bayesian prediction

$$P(Y_{T+1}|D) = \int P(Y_{T+1}|\theta,M,D)P(\theta|M,D)P(M|D)\mathrm{d}\theta\mathrm{d}M, \quad (4)$$

integrates the uncertainty in the model structure and parameters.

2.2 MapReduce

As the quantity of manufacturing process data is getting bigger and bigger, traditional machine learning and data mining methods for BN learning can no longer output valid results within reasonable time bound. As an important tool for big data analytics, MapReduce offers possible solutions for BN learning under the circumstance of large scale data.

MapReduce is a programming model and an associated implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks [10]. Map/reduce is applicable in a wide range of use cases. It is organized as a "map" function which transform a piece of data into some number of key/value pairs. Each of these elements will then be sorted by their key and reach to the same node, while a "reduce" function is use to merge the values (of the same key) into a single result. Ho [19] further explained the basic principles and general workflow of MapReduce in detail as presented in Figure 1.

Some exploratory research on integrating BN and MapReduce from the perspective of computing theory are discussed [20–23], most of which are published in conferences without practical application. This paper emphasizes on factor modelling and ERV based key reaction chain analysis in quality control area.

3 Proposed Framework Integrating Big Data Analytics and BN Approach

A framework for manufacturing quality control integrating big data analytics and BN is proposed as presented in Figure 2. The framework consists of four layers from bottom to up. The bottom layer is data acquisition. The second layer is a distributed file storage system. The third layer is the core layer of the whole system, which is the parallel computing framework integrating BN and MapReduce. The top layer mainly integrates user interface and system interface, in order to define personalized data needs, formulate specific algorithm rules and realize data visualization. A feedback loop is added from the top layer to the bottom layer based on analysis results and user needs.

(1) Bottom layer: data acquisition. Manufacturing systems contains multiple factors [24], including humans, equipment, materials, processes, environments, measurements, etc. Any of them could influence the manufacturing quality. Data can be obtained from sensors, manual collection and generated from the calculation of information systems.
(2) Second layer: distributed file storage system. Different types of data are stored in the HDFS. On one hand, data security is guaranteed through data redundancy. On the other hand, big volume of data can be shared though exchange and interaction between different nodes.
(3) Third layer: the parallel computing framework which integrates BN and MapReduce. Data are divided into splits and released to independent memories and processors though the Map task. Processors execute the Reduce task with the pre-set rules, which are referred to BN learning algorithms here.
(4) Top layer: input and output interfaces both for users and systems. Users can customize the final data type and computing rules. Meanwhile, visualized and dynamic computing results can be presented to users, including structured correlation networks and probability distribution tables, which offer supporting information for further decision-making and manufacturing system adjusting.

Research of the third layer is the core work of this paper. First, BN based quality data analysis process is studied. Then, under the circumstance of big data, supporting MapReduce is applied in the learning process of BN structure and parameters.

4 Process of BN Based Manufacturing Quality Data Analysis

Due to advantages in dealing with uncertain and dynamic problems, BN is getting widely used in intelligent decision-making, data fusion, pattern recognition, medical diagnosis, data mining and other areas. Because of the development of BN recent years, taking time series into account, application fields and problem solving ability got extended and improved. Especially in data mining, BN is used for classification, regression analysis, causal reasoning, uncertain knowledge representation, clustering pattern

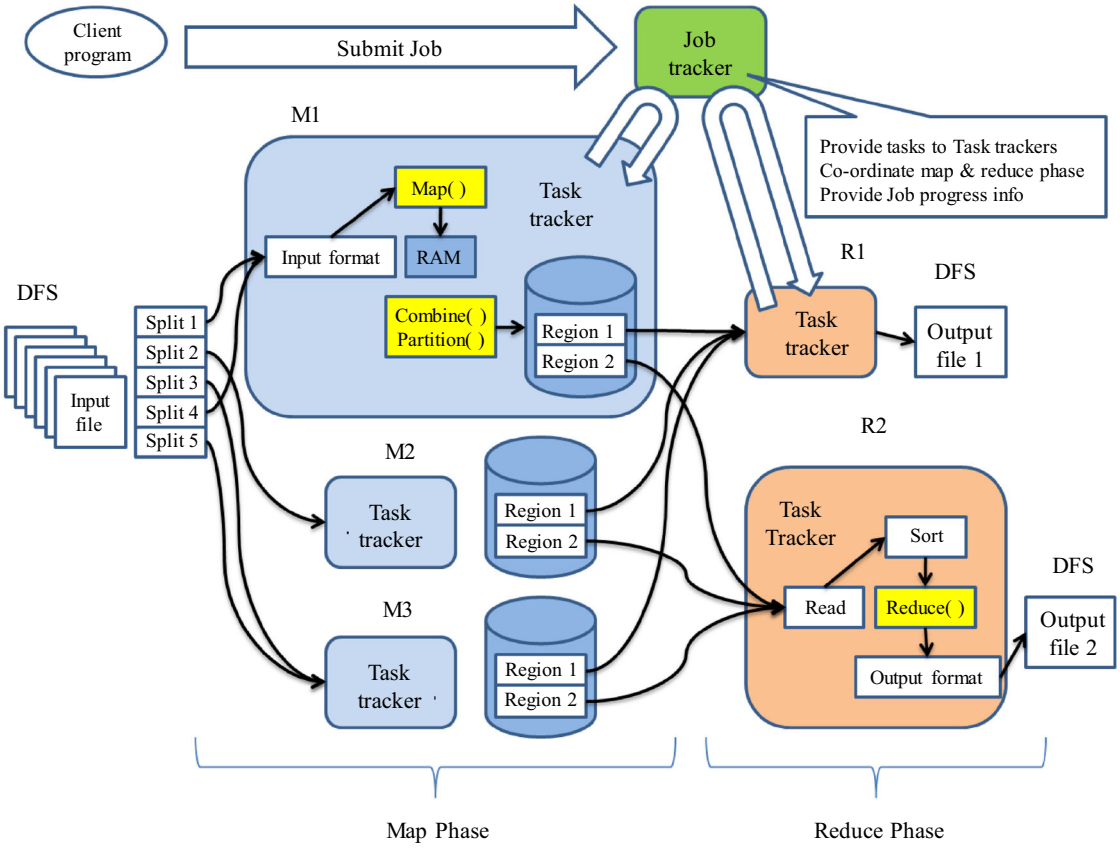

Figure 1 Basic principle and general workflow of MapReduce [19]

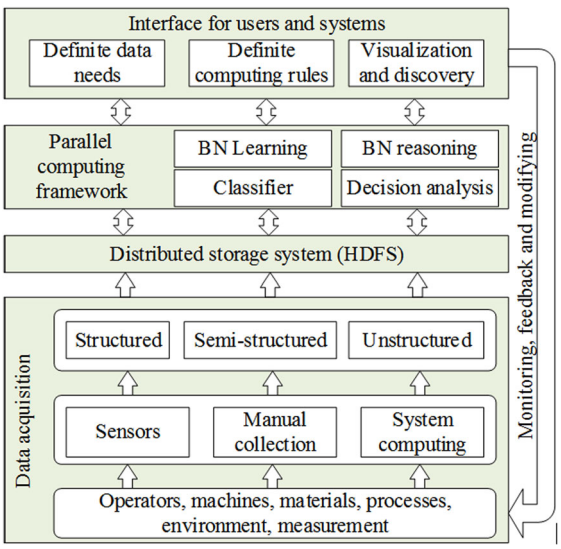

Figure 2 Proposed framework for quality control based on big data analytics and BN

discovery, and so on. In manufacturing industry, BN has been applied to reliability analysis and other areas. Therefore, BN also can be a useful tool for manufacturing process quality problems analysis and control.

The key of BN problem solving is the network construction, which contains two basic components, namely the network structure I and conditional probability table (CPT) E. Generally, BN structure can be constructed according to expertise experiences, or learned from massive associated data. The former method relies on experts' subjective and probably one-sided knowledge and experience, with low credibility. But the latter method offers a way of obtaining implicit uncertain knowledge in massive data. The BN network structure got from this approach is accurate and stable, which is approved widely by researchers.

In manufacturing process, without considering the requirements for computing ability and speed, a basic workflow for manufacturing quality control with BN from the view of logic is proposed as depicted in Figure 3.

(1) First, all the possible factors influencing manufacturing quality should be listed, as alternative nodes for BN construction. To improve the efficiency of structure analysis, tools of functional analysis and substance-field models could be applied.
(2) Collect all the possible data related to all the factors. Feature subset selection (FSS) method are used to

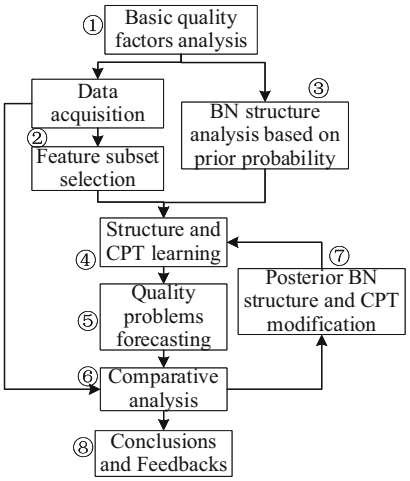

Figure 3 Manufacturing quality problem analysis and control process based on BN

eliminate those unrelated or low related factors, to simplify the model structure and reduce computing complexity.

(3) Establish the preliminary BN structure and probability distribution table with expert knowledge and experience.

(4) Through machine learning and data mining from the massive data, rebuild and modify the BN model acquired in step (3).

(5) According to the input conditions of all the father nodes, the final obtained BN is applied to forecast the outcomes of target child nodes.

(6) Compare the forecasting results with those data acquired from reality in the following manufacturing processes, and analyse classification accuracy.

(7) With the analysis results in step (6) and based on posterior probability to modify the BN structure and CPT. K2 score [25] and EM algorithm are applied for iterative learning, computing and optimizing for the BN model.

(8) Then the BN model with stable structure and CPT is used for root cause and key reaction chain analysis, offering decisions for quality problem decomposition and solution proposing.

5 MapReduce Enhanced BN Learning and Key Reaction Chain Identification

5.1 K2 Score Based BN Structure Learning

The learning of BN structure is an NP hard problem. Here, MapReduce and K2 score [25] are introduced to accelerate the computing process. K2 is a method of searching and score for BN structure learning. The basic thought of this method is to list the most possible several candidate network structures, and score them with specific algorithm to measure their coincidence degree. Suppose $N = \{X_1, X_2, \ldots, X_n\}$, in which the value range of X_i are$\{ x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$, $r_i \geq 2, i = 1, \cdots, n$. G represents a structure of N. Then the highest scored network model G^* with highest posteriori probability can be figured out:

$$G^* = \arg\max P(G|D), \tag{5}$$

where D is the original data sheet. The values of $P(G|D)$ can be calculated with the following Eq. (6). Detailed proof process can be found in [25].

$$P(G|D) = g(i, \pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!. \tag{6}$$

As data record items are independent from each other, data can be split into segments and assigned to different nodes with MapReduce operations for computing. The network model G^* with the highest $K2$ score will be chosen as the final result of the BN structure, candidate for further parameter learning and analysis.

5.2 BN Parameter Learning Based on EM Algorithm

Parameter (CPT) learning, is another key issue in solving BN problems. Expectation Maximization (EM) is an acknowledged iterative algorithm for learning statistical models, including BNs from data with missing values or latent variables [26, 27]. The integration of MapReduce in EM algorithm can accelerate the learning process greatly. Base on the work of [20, 21], CPT learning of BN based on MapReduce can be decomposed into two steps.

E-step: Each mapper takes as input BN structure β, current estimate of parameters θ^t, a junction tree (JT), decomposition of the BN structure T, and data sheet D. It runs the E-step on its input records and accumulates pseudo-counts $\bar{M}[x_i, \pi_{x_i}]$ for all child-parent combinations in a hash map. Once the mapper processes of all records assigned to it, it emits an intermediate key-value pair for each hash map entry. The emitted key contains state assignments to parents π_{x_i} of the node X_i, whereas the value represents the child variable assignment x_i appended with the soft counts $\bar{M}[x_i, \pi_{x_i}]$ for this entry. This intermediate key makes sure that all variables with the same parents are grouped and processed in the same reduce task.

M-Step: Each reduce method performs the M-step for families with the same parent assignment: it iterates through all the values with the same key, parses the value, and fills a hash map, in which keys correspond to child-parent combinations and their states, and values correspond

to the soft counts. Values are summed up to obtain the parent count. Finally, each reduce function emits an output key-value pair for each hash map entry. The output key is of the form; the output value represents a newly estimated parameter $\theta^{t+1}_{x_i|\pi_{x_i}}$.

5.3 ERV Based Key Reaction Chain Identification

With previous two steps, a quality assessment BN model with CPT can be obtained. Basic relationships between factor nodes and the objective can be observed by the visualized BN structure. However, terminal nodes of BN models are just the final manifestation of factor reaction chains. Elimination of some obvious hazard factors may just be a stopgap which would cost a lot and leave hidden recurrence problems. To find out those key root causes and map out those reaction chains, a method of entropy reduction value (ERV) based sensitivity analysis is applied. ERV provides a means of evaluating the sensitivity of each node to their son nodes [28], which has been used for risk analysis [29]. Detailed calculation process is presented as follows.

Suppose X and Y are two random variables, and x, y are their given values. Then the mutual information ERV of X and Y can be defined as:

$$I(X|Y) = H(X) - H(X|Y), \tag{7}$$

in which

$$H(X) = -\sum_{x\in X} p(x)\log p(x), \tag{8}$$

and

$$H(X|Y) = -\sum_{y\in Y}\sum_{x\in X} p(y)p(x|y)\log p(x|y). \tag{9}$$

Detailed proof and derivation procedures can be find in the work of Li, et al. [30] The measure in Eq. (7) represents the difference of the a priori and a posteriori entropies of X, i.e., the reduction in uncertainty about X by knowing Y. It is noted that $I(X|Y)$ is the information in X about Y, and a high degree of dissimilarity indicates more information X carries about Y.

6 Case Study: Management of Construction Precision in Shipbuilding Industry

6.1 Background of the Case Study

Recent years, with the promotion of information technology and intelligent manufacturing, as well as the increasing needs of domestic and international market, shipbuilding industry get rapid development. Products in shipbuilding industry are characterized with large size, complex structure, high cost and long manufacturing period. Data amount in the whole construction period are far beyond the control scope of traditional database technology.

IBM has helped Daewoo of South Korea in dealing with the big data problems, including enterprise resource planning and database management, which involves more than 255 TB data. In 2013 November, Mitsubishi Heavy Industries also declared their collaboration with NEC Corporation in the development of an "Energy Demand Forecast System for Ships" applying NEC's big data analysis technologies to achieve energy savings during ship navigation. However, the applications of big data in shipbuilding industry are limited in data storage and operation data analysis. Applications in manufacturing processes, especially in quality control, has not been studied and reported.

6.2 Shipbuilding Construction Precision Management

According to design drawings, steel ships and marine equipment are constructed with processes of lofting, marking off, machining, assembling, welding, hoisting, and so on. Manufacturing process has various factors that may influence the construction accuracy. Physical parts, components, segments of the hull would inevitably have shape and size deviation from the loft model, not mention the design model. To control these deviations within the scope of standard requirements, shipyards usually add allowance to hull parts when lofting, and cut off redundant margins before assembling and welding. This will inevitably lead to considerable on-site trimming workload. Almost all the trimming work has to be finished by manual operation. The consumption of working time accounts for 1/4 of the total hull construction work. To reduce those unnecessary wastes, methods of using compensation instead of allowance was developed to control the deviation. The compensation method needs almost no trimming work, which is based on acquisition and mathematical statistics of massive data from production and measurement practice. Figure 4 and Table 1 provide a classical case in hull segment manufacturing precision management.

Some software, e.g., Samin EcoMES® and Haily DACS®, are now widely used in shipbuilding accuracy control. However, most of them run offline stand-alone on single computers. The operation system fundamentally restricts the processing and analysis abilities. Meanwhile, correlation analysis of factors is based on traditional statistics methods, only one or two level mapping relationships can be analysed. When comes to deeper layer mapping relationships, it lacks efficient methods and tools.

Constructions of ships and marine engineering equipment are generally huge projects with massive data items

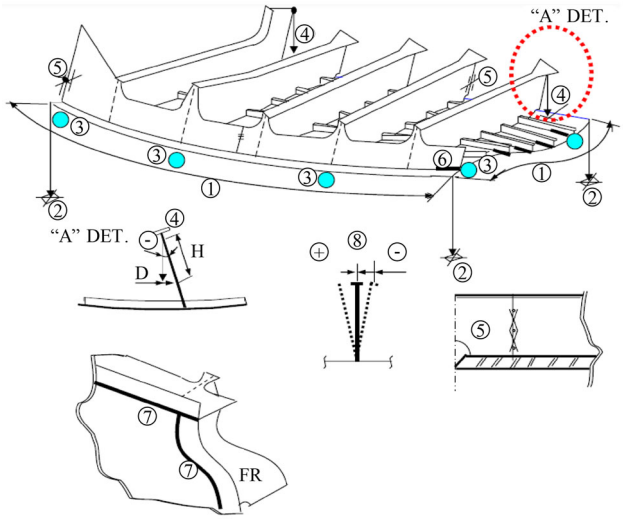

Figure 4 Control points for hull segment precision management. Note: Explanation of the numbered contents are shown in Table 1

Table 1 Benchmarks and limits of hull segment manufacturing precision management

No.	Checking positions	Benchmarks /mm	Limits /mm
1	Length (actual)	±3	±6
2	Width (height)	±3	±5
3	Level	±4	±6
4	Verticality	±4	±6
5	Telos deviation	±3	±5
6	Welding reservation	Standard	
7	End back burning	0 ~ +3	0 ~ +6
8	Vertical angle	±4	±6

and types accumulation. Especially when multi projects run simultaneously, construction accuracy control challenges the computing and processing ability of enterprise information systems.

In the proposed BN and big data analytics integrated method, distributed storage and parallel computing are applied to the solving of BN based data mining, which offers an effective strong support to accuracy control of shipbuilding. Coinciding to the thought of compensation, the latest analysis method will greatly reduce extra on-site trimming work, which will further shorten construction period and cut down costs.

6.3 Data Sources and Characters for Shipbuilding Construction Accuracy Control

Shipbuilding precision is an interaction result of multi factors. Data acquired about operators, machines, materials, processes, environments and measurements, will all have great potential values. Detailed data of each factor category is as shown as follows:

(1) Data of operators: age, gender, work experiences, skill level, past operating records, etc.
(2) Data of machines: application, equipment model, rated parameters, tool types, running time, historical status, failure and maintenance records, etc.
(3) Data of materials: material type, model, ingredient, size, weight, suppliers, storage method, transport mode, etc.
(4) Manufacturing processes: baseline positioning methods, welding parameters, deformation compensation method, cutting process of sheet metal, splice sequences, etc.
(5) Measurement data: actual size, size deviation, margin size, deformation, breakage, welding defects, etc. Meanwhile, measuring tools, methods and locus should be recorded.
(6) Environment parameters: temperature, humidity, noise, vibration, etc. Past records, real-time data and future forecasts are all needed.

Those data are acquired from different ways and have different data structure. There are numbers, texts, images and even audios. Data may be types of continuous, discrete or Boolean. General approaches may find difficulty in processing all the information. Big data analytics offers a powerful tool in dealing with this problem. To build the BN model, different kinds of data are dispersed based on different principles and calculate statistical probabilities of discrete data segments. In this way, differentiate data sources and data types can be normalized with probability and correlated in the same BN platform based on MapReduce. In the BN model for shipbuilding construction accuracy control, the authors mainly focus on the quality analysis of discrete process stages and the data are collected and correlated along the process flow line.

6.4 Hull Segment Manufacturing Precision Control with the Proposed Approach

As a significant process of shipbuilding, manufacturing precision of hull segments influences the overall construction accuracy greatly. The manufacturing of hull segments is a complex combination of welding operations. Failure control in deformation and defects may lead to huge hidden risks to the safety and reliability. Size and shape deformations can be measured through large coordinate measuring machines. Welding defects now can be obtained through image recognition and ultrasonic testing. Information of every weld bead will be numbered, marked, classified and stored. Data of other manufacturing quality related factors are collected in the similar way. Data items

collected include designer experience, worker skill, guidance documents, welding difficulty, equipment level, process environment, process flow, operation failure, process management, deformation, change & rework, etc. Historical data items should be considered. About 1.3 million data items (8.9 TB) related about welding process in hull segment manufacturing are collected for analysis. Data items are pre-processed into discrete numeric or semantic values according to shipbuilding precision management and control operation guidance. The analysis and control objectives of shipbuilding precision data are acquired by two steps as shown in Figure 5. First, by onsite measurement with total stations, relative coordinates of every control point are recorded promptly with a dedicated PDA. Then, the data are imported to the database of Haily DACS® for numerical and geometric fitting with the digital design model. Manufacturing deviation of very control point and an average can be acquired for manufacturing quality assessment.

The Hadoop cluster is set up on two Intel Xeon server machines with 40 cores CPU, 2.27 GHz processor, 128 GB memory and 100 Mb/s Ethernet LAN. Oracle Virtual Box is installed to configure sixteen VMs on the server. Each VM was assigned with 4 CPU cores, 8 GB RAM and 150 GB hard disk storage. The Hadoop-1.2.1 version was installed, and one VM is configured as Name Node and the remaining 15 VMs are configured as Data Nodes. Then next follows the analysis process. Data files are imported into HDFS to reduce the pressure of memories. Data files are divided and stored into distributed storage and physical addresses. Then, users specify analysis objectives, file inputs, BN learning evaluation criteria and stop conditions. MapReduce deploys multi-workers to carry out Map and Reduce tasks. Distributed computing nodes perform computing process according to $K2$ score and EM algorithm presented in Sections 5.1 and 5.2 for DBN structure and parameter learning and modification.

Last, a complete structured hull segment manufacturing precision BN model with CPT is returned as a description data file. To evaluate the efficiency of proposed approach, a computing ability test with different amount of computing nodes and data records is conducted as depicted in Figure 6. It shows that the introduction of MapReduce in BN learning can significantly reduce the computing time. As for the calculation accuracy which is concerned the most, it turned out that the BN structures and CPTs under different number of computing nodes are almost the same.

The result of the output file is imported to a BN analysis software GeNIe® for graphical modelling, which is presented in Figure 7. Network nodes including DE (designer experience), WS (worker skill), GD (guidance documents), WD (welding difficulty), EL (equipment level), PE (process environment), PF (process flow), OF (operation failure), PM (process management), DM (deformation), CR (change & rework) and MP (manufacturing precision). Then ERVs of every father node to their son nodes are calculated and marked on the directed arrows following the ERV method presented in Section 5. The values of ERVs are represented by the width of arrow lines. Following the red lines which are recognized with relatively high ERVs, four key reaction chains can be mapped out, including $DE \rightarrow WD \rightarrow PM \rightarrow CR \rightarrow MP$, $DE \rightarrow WD \rightarrow OF \rightarrow (CR) \rightarrow MP$, $DE \rightarrow WD \rightarrow PF \rightarrow (DM) \rightarrow MP$, and $PE \rightarrow (PF) \rightarrow DM \rightarrow MP$.

With the presented BN model, root causes to manufacturing precision can be figured out. In this case, the designer factor influences the final manufacturing precision significantly, which is usually be ignored because it is separated with the manufacturing process. And in the same

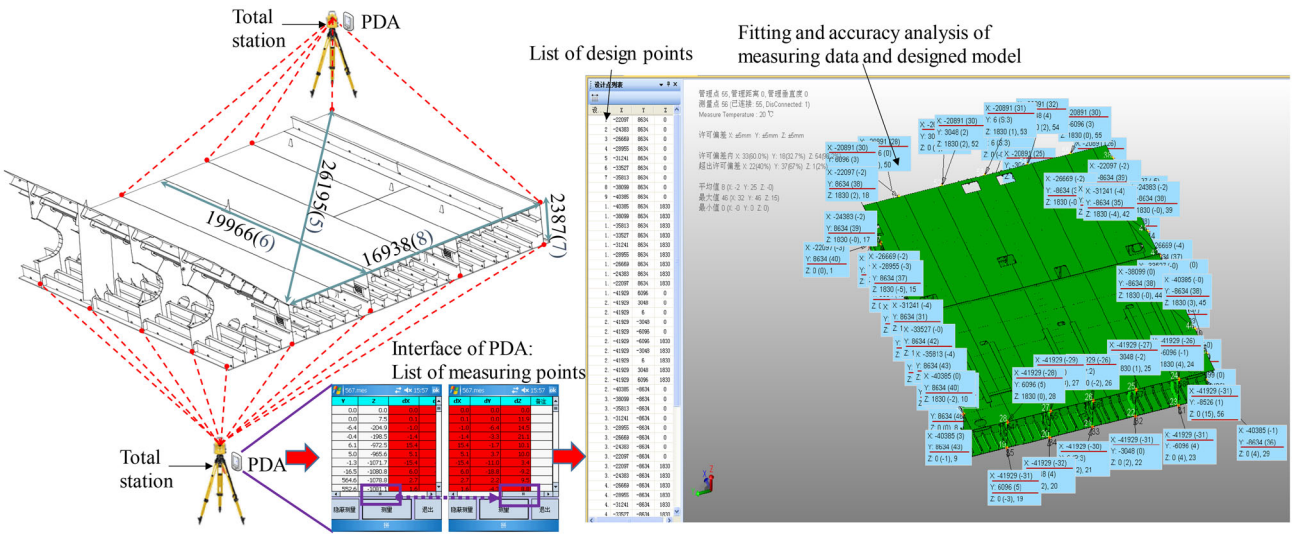

Figure 5 Process of shipbuilding manufacturing accuracy analysis

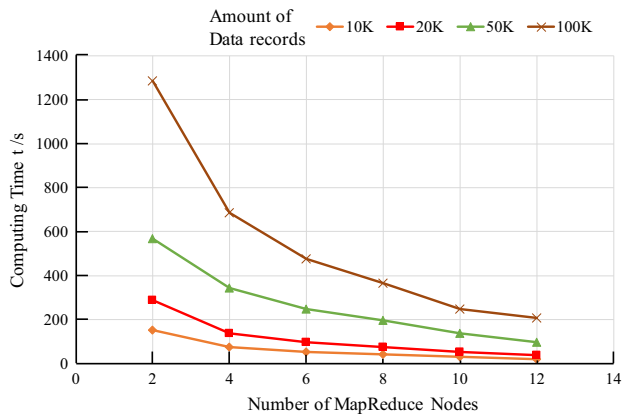

Figure 6 Computing ability test with different amount of nodes and data records

time, key reaction chains can offer comprehensive decision support for decision makers. Those intersections of key reaction chains are identified as key nodes, like *PF*, *OF*, *DM* and *CR*. The volatility of those key nodes will influence the CPT of BN model greatly, which should be taken more seriously in the precision control process.

Of course, hull segment manufacturing precision analysis is only a key part of the whole construction accuracy control process. With the proposed quality analysis and control model, an overall BN model for shipbuilding construction accuracy control could be further established. Related data can be obtained continuously. Integrating priori knowledge and posterior experience, the model can be modified iteratively, so that the accuracy, stability and reliability for system diagnosis, prediction and decision can be improved.

7 Conclusions

(1) A four-layer framework integrating BN and big data analytics is proposed for manufacturing quality control, which has both advantages of BN in dealing with uncertain problems and MapReduce in processing large scale data.

(2) Detailed manufacturing quality control processes with BN and Specific approaches for accelerating BN learning process with MapReduce are developed.

(3) ERV method is introduced for root cause identification and key reaction chain analysis, which offers quantized and visualized decision support for decision makers.

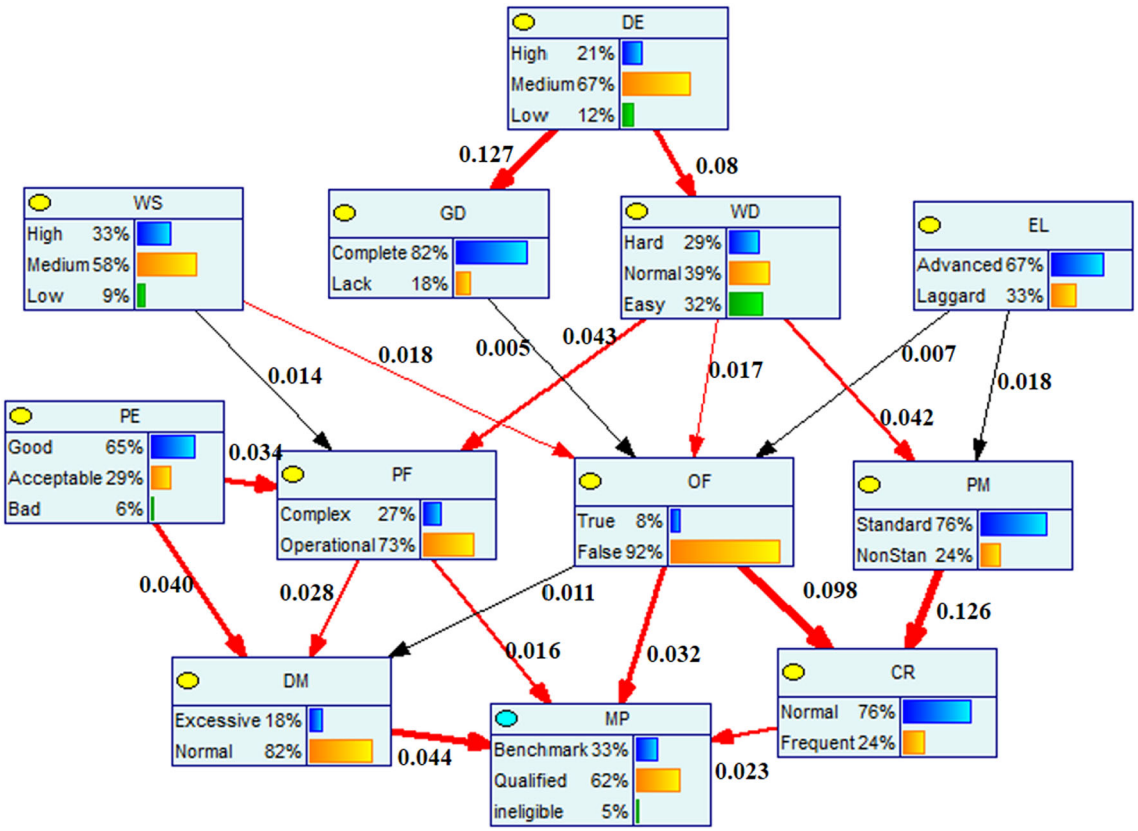

Figure 7 BN model for hull segment manufacturing precision analysis

(4) The case study in shipbuilding construction accuracy control shows that computing speed acceleration is almost directly proportional to the increase of computing nodes and four key reaction chains are identified, which verifies the efficiency and feasibility of the approach, and reveals its capability and further application potentials.

(5) Future research will focus on the development of a mature information system to support the automatic integration of executing procedures. Detailed technique solutions will also be researched for modification and improvement.

References

1. H Kagermann, J Helbig, A Hellinger, et al. *Recommendations for Implementing the strategic initiative INDUSTRIE 4.0: securing the future of German manufacturing industry; final report of the Industrie 4.0 working group*. Berlin: Acatech, 2013.
2. S Q Feng, H Terasaki, H Komizo, et al. Development of evaluation technique of GMAW welding quality based on statistical analysis. *Chinese Journal of Mechanical Engineering*, 2014, 27(06): 1257–1263.
3. J Manyika, M Chui, B Brown, et al., *Big data: The next frontier for innovation, competition and productivity*. San Francisco: McKinsey Global Institute, 2011.
4. D Wong. *Data is the Next Frontier, Analytics the New Tool*. London: Big Innovation Centre, 2012.
5. C F Chien, C W Liu, S C Chuang. Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement. *International Journal of Production Research*, 2015, 29(9): 1–13.
6. R Y Zhong, C Xu, C Chen, et al. Big Data Analytics for Physical Internet-based intelligent manufacturing shop floors. *International Journal of Production Research*, 2015: 1–12.
7. F Chang, J Dean, S Ghemawat, et al. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 2008, 26(2): 1–26.
8. S Ghemawat, H Gobioff, S T Leung. The Google file system. *ACM SIGOPS Operating Systems Review*, 2003, 37(5): 29–43.
9. K Shvachko, H Kuang, S Radia, et al. The hadoop distributed file system. *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Lake Tahoe, NV, USA, May 3–7, 2010. Washington DC: IEEE Computer Society, 2010: 1–10.
10. J Dean, S Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 2008, 51(1): 107–113.
11. J L Deng. Introduction to grey system theory. *The Journal of grey system*, 1989, 1(1): 1–24.
12. Z Pawlak. Rough sets. *International Journal of Computer & Information Sciences*, 1982, 11(5): 341–356.
13. G Z Diao, L P Zhao, Y Y Yao. A System Framework of Dynamic Coupled Quality Control Based on Complex Network. *Advanced Materials Research*, 2013, 711: 773–778.
14. F V Jensen. *An introduction to Bayesian networks*. London: UCL Press, 1996.
15. K P Murphy. *Dynamic bayesian networks*. California: UC Berkeley, 2002.
16. Z Ghahramani. Learning dynamic Bayesian networks//*Adaptive processing of sequences and data structures*. Berlin: Springer Berlin Heidelberg, 1998: 168–197.
17. Z J Yang, Y N Kan, F Chen, et al. Bayesian reliability modeling and assessment solution for NC machine tools under small-sample data. *Chinese Journal of Mechanical Engineering*, 2015, 28(6): 1229–1239.
18. N Friedman, I Nachman, D Peér. Learning bayesian network structure from massive datasets: the sparse candidate algorithm. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Stockholm, Sweden, July 30–August 1, 1999. Burlington: Morgan Kaufmann Publishers Inc., 1999: 206–215.
19. R Ho. *Hadoop Map/Reduce Implementation*. http://horicky.blogspot.sg/2008/11/hadoop-mapreduce-implementation.html. 2008-12-05/2016-05-19.
20. A Basak, I Brinster, X Ma, et al. Accelerating Bayesian network parameter learning using Hadoop and MapReduce. *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*. Beijing, China, August 12–16, 2012. New York: ACM, 2012: 101–108.
21. A Basak, I Brinster, O J Mengshoel. Mapreduce for Bayesian network parameter learning using the EM algorithm. *Proc of Big Learning: Algorithms, Systems and Tools*, 2012, 12: 1–6.
22. W Chen, T Wang, D Yang, et al. Massively parallel learning of Bayesian networks with MapReduce for factor relationship analysis. *2013 International Joint Conference on Neural Networks (IJCNN)*. Dallas, TX, USA, Aug 4-9, 2013. New York: IEEE, 2013: 1–5.
23. Q Fang, K Yue, X Fu, et al. A mapreduce-based method for learning bayesian network from massive data. *15th Asia-Pacific Web Conference on Web Technologies and Applications*. Sydney, NSW, Australia, April 4–6, 2013. Berlin: Springer Berlin Heidelberg, 2013: 697–708.
24. X Xi, X Z Wu, Y L Wu, et al. Modeling and analysis of mechanical Quality factor of the resonator for cylinder vibratory gyroscope. *Chinese Journal of Mechanical Engineering*, 2016, 30(1): 180–189.
25. G F Cooper, E Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, 1992, 9(4): 309–347.
26. A P Dempster, N M Laird, D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 1977, 39(1): 1–38.
27. N Friedman, K Murphy, S Russell. Learning the structure of dynamic probabilistic networks. *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Madison, WI, USA, July 24-26, 1998. San Francisco: Morgan Kaufmann Publishers Inc., 1998: 139–147.
28. G S Hamilton, F Fielding, A W Chiffings, et al. Investigating the Use of a Bayesian Network to Model the Risk of Lyngbya majuscula Bloom Initiation in Deception Bay, Queensland, Australia. *Human and Ecological Risk Assessment: An International Journal*, 2007, 13(6): 1271–1287.
29. M K Zheng, X G Ming, M Li, et al. A framework for Industrial Product–Service Systems risk management. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 2015, 229(6): 501–516.
30. M Li, L Wang. Feature fatigue analysis in product development using Bayesian networks. *Expert Systems with Applications*, 2011, 38(8): 10631–10637.

Mao-Kuan Zheng, born in 1989, is currently a PhD candidate at *School of Mechanical Engineering, Shanghai Jiao Tong University, China*. He received his bachelor degree from *China University of Geosciences, China*, in 2011. His research interests include intelligent manufacturing and product service systems. Tel: +86-21-34206528; E-mail: zhengmaokuan@163.com

Xin-Guo Ming, born in 1966, is currently a professor at *School of Mechanical Engineering, Shanghai Jiao Tong University, China*. He received his PhD degree on Mechanical Engineering from *Shanghai Jiao Tong University, China*, in 1995. His research interests include product innovation management, intelligent manufacturing and product service systems. E-mail: xgming@sjtu.edu.cn

Xian-Yu Zhang, born in 1987, is currently a PhD candidate at *School of Mechanical Engineering, Shanghai Jiao Tong University, China.* His research interests include smart manufacturing and personalized customization. E-mail: 513615330@qq.com

Guo-Ming Li, born in 1990, is currently a master candidate at *School of Mechanical Engineering, Shanghai Jiao Tong University, China.* His research interests include smart manufacturing and big data analytics. E-mail: xiaohebei1990@163.com