Notes on Data-driven System Approaches

XU Jian-Xin[1] HOU Zhong-Sheng[2]

Abstract In this paper, we present several considerations centered around the data-driven system approaches. We briefly explore three main issues: the evolving relationship between off-line and on-line data processing methods, the complementary relationship between the data-driven and model-based methods, and the perspectives of data-driven system approaches. Instead of offering solutions to data-driven system problems, which is impossible at the present level of knowledge and research, in this article we aim at categorizing and classifying open problems, exploring possible directions that may offer alternatives or potentials for the four key fields of interests: control, decision making, scheduling, and fault diagnosis.

Key words Data-driven, principal relations, problem classification, research directions

In large scale and complex industrial processes such as in oil refinery plants, traffic and communication networks, power grids, aeronautics and astronautics, a most prominent feature is the presence of vast volume of data accompanied by the lack of an effective process physical model that can support control, fault diagnosis, scheduling, and decision making. Here, the large scale refers to the scale level from a few dozen to a few thousand variables or nodes, and the complexity arises from heterogeneous information sources, multi-modal signals, high non-linearities, strong interactions among variables or system states, involvement of human activities, and mixture of tasks from the four key fields in one integrated process. Often, the direct consequence of the large scale and complexity is the wide spread system uncertainties that prevent the problem solving from using a physical model-based approach.

The main objective of this paper is to seek possible alternatives coined by data-driven system approaches in general, and off-line data processing technology in particular. Specifically, we focus on the four key fields of interests and look into other areas that would provide non-conventional possible solutions, as shown in Fig. 1.

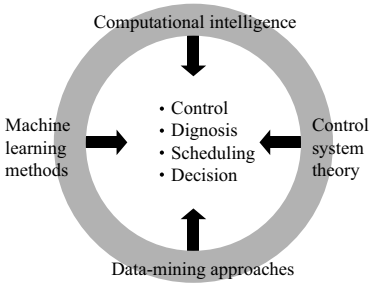

Fig. 1 The four key fields of interests are control, decision making, scheduling, and fault diagnosis. The other four areas of research considered here are computational intelligence, data-mining, control system theory, and machine learning methods.

Though overlapping in certain areas of applications, the four areas of research cover almost all methods or algorithms that are used in solving problems associated with various kinds of data, on-line or off-line, homogeneous or heterogeneous, deterministic or stochastic, etc. Fig. 2 summarizes briefly the field of data-mining[1−3]. Fig. 3 shows the field outline of machine learning[4−6]. Fig. 4 presents several main subfields and approaches in computational intelligence[7−9]. Fig. 5 gives a glance at the field of the system control[10−18].

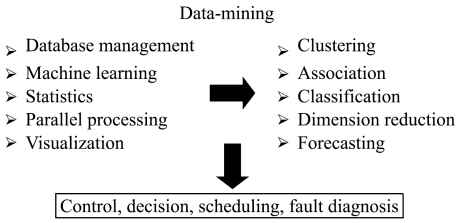

Fig. 2 The outline of data-mining approaches and applications that could be relevant to data-driven systems in control, decision making, scheduling, and fault diagnosis

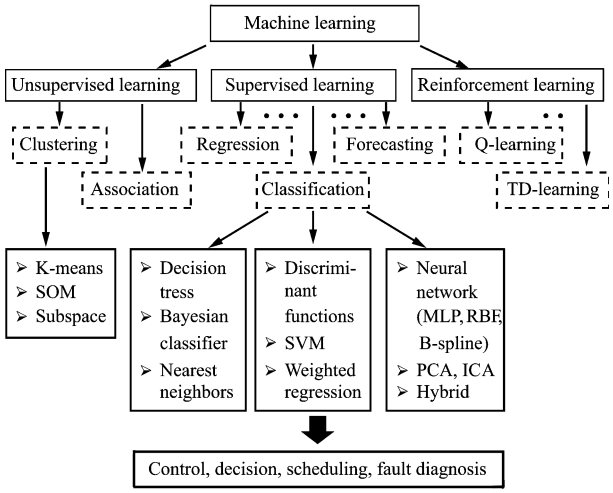

Fig. 3 The outline of machine learning approaches that could be relevant to data-driven systems in control, decision making, scheduling, and fault diagnosis

Data-driven approaches have been widely applied to solve industrial and real-life problems encountered in the four key fields, including control engineering[19−21], instrumentation and measurement[22], computer security[23−24], power grid and systems[25−26], intelligent transportation systems and vehicles[27], aerospace[28], circuit design and integrated circuit scheduling[29], diagnosis and planning in medicine and rehabilitation[30−31], Internet and web[32], electrical drives[33−34], process industry supervision[35], manufacturing industries such as semiconductor industry[36].

Received December 17, 2008; in revised form March 9, 2009
Supported by State Key Program of National Natural Science Foundation of China (60834001) and National Natural Science Foundation of China (60774022)
1. Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576, Singapore 2. Advanced Control Systems Laboratory, School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, P. R. China
DOI: 10.3724/SP.J.1004.2009.00668

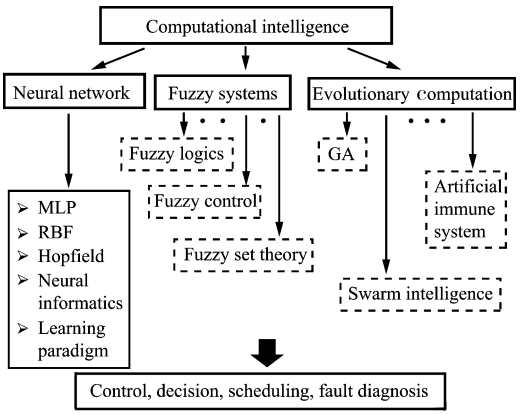

Fig. 4 The outline of computational intelligence approaches that could be relevant to data-driven systems in control, decision making, scheduling, and fault diagnosis

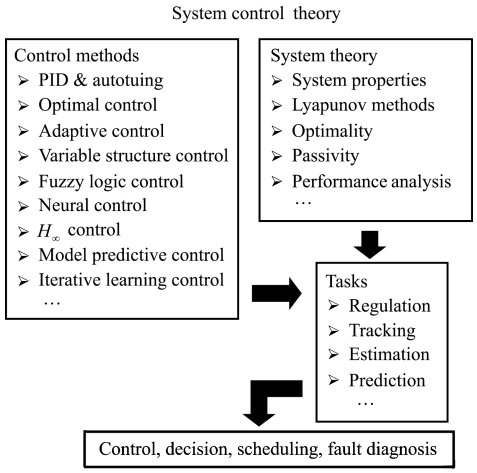

Fig. 5 The outline of system control approaches that could be relevant to data-driven systems in control, decision making, scheduling, and fault diagnosis

Machine learning and computational intelligence approaches are also widely used to address the problems in the four key fields of control[37−39], decision making[40−41], scheduling[42−43], and fault diagnosis[44−46].

It is worthwhile pointing out that the possible combination of data mining and computational intelligence methods has been also explored, for instance integrating data mining with neural network[47], fuzzy logics[48], and evolutionary algorithms[49].

The recent advances in data-driven system approaches, as evidenced by preceding discussions and published reports on the applications and theoretical research, provide a clear picture on the relationship in between the four key fields of interests — control, decision making, scheduling, and fault diagnosis, and the other four areas of research — computational intelligence, data-mining, control system theory, and machine learning methods, as described in Fig. 1.

1 The relationship between on-line and off-line methods

The central idea of data-driven system approaches is to use off-line data in some way to enhance on-line data-driven system approaches when the latter cannot meet the performance requirements. As such, an important issue is to make it clear the major differences between the off-line data and the on-line data.

Note that a vast amount of off-line data does not warrant the presence or acquisition of new system information beyond the existing ones from the on-line data. This issue in a sense is analogous to the richness condition in system identification or the persistent excitation condition in adaptive control. Thus, the first expected characteristic of off-line data should be the sufficiency, which is illustrated in Fig. 6, in which the rectangle represents a sample space with the necessary dimensions and size. The sample space is a collection of off-line samples shown as dots in the figure. What we expect from off-line data is the abundance of samples that evenly spread over the range of interests in the sample space. Here, a sample could be an action, a decision, a feature, an observation, etc., which is a sequence of off-line data recorded along temporal or spatial coordinates.

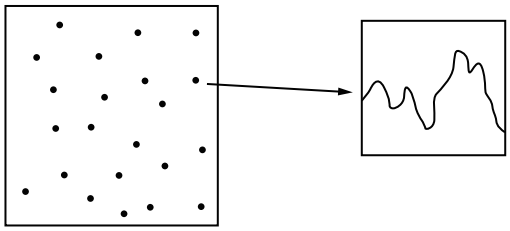

Fig. 6 The sufficiency of the sample space implies the ergodicity. Graphically, we expect that the samples could be evenly distributed in the sample space.

Another important characteristic of the off-line data is the completeness. The off-line data sequence in a sample should be complete or dense almost everywhere. Often in industry, the data collection process is incomplete and data dropout occurs. For instance, in the monitoring process of traffic flow in a city, surveillance devices such as cameras or loop detectors are rather sensitive to environmental conditions such as lighting condition, vehicle size and speed, or the reliability of devices which may lead to severe data missing at some period and at some location. Several issues need to be addressed for the completeness of off-line data: the necessary level of completeness associated with a given task, the evaluation on the completeness, the statistic completeness, and active compensation for the completeness. Although the completeness is difficult to achieve in a single sample, it is possible to achieve the statistic completeness in an ensemble sense. The active compensation is to purposely reproduce a sample or to generate a new sample that can cover the missing parts from historical records. To achieve the completeness of off-line data, it is necessary to perform appropriate posterior analysis, pattern recognition and classification for off-line data, decide if a deficiency exists, and seek a remedy if it is possible.

As shown in Fig. 7, on-line data is the reading from a moving window of appropriate size. In contrast to historical data obtained off-line, on-line data captures the latest information, reflect varying characteristics of a process or an event, and are used for real-time systems. Because on-line data consists of only a portion of a sample, it suffers from local information, data loss, and measurement bias.

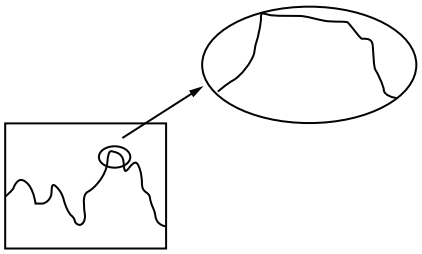

Fig. 7 On-line data acquisition within a moving window

Clearly, off-line data would contain much richer system information than on-line data, which motivates researchers to explore suitable data processing methods or algorithms so that off-line data can be used as though it is on-line. A question is whether we need to develop new methods and algorithms for this goal. There are cases where we can directly use historical data in real time with little changes or modifications, such as in iterative learning control. In other cases, however, off-line data cannot be incorporated directly, instead it should be incorporated indirectly in the form of a data-driven model as we will further explore later, or in an application-specific form such as a rule base, a lookup table, and a characteristic model, by means of off-line data processing methods.

It is worth to point out that the available pool of off-line data processing methods is far ample than that of on-line. In fact, most theoretical methods and successful applications in the areas of data mining, machine learning, and computational intelligence are off-line in nature. If possible, we should make full use of existing off-line methods that have been proven to be effective through the vast academic exploitations and real-life applications.

There are two reasons that justify the extension of off-line data processing methods in solving on-line problems associated with the four key fields of interests. First, many on-line problems from industry are temporally multi-scale, for instance a mixture of short-term and long-term goals and actions in control or decision making. As shown in Fig. 8, the short interval only allows the use of simple real-time data-processing algorithms, whereas the large interval allows the use of computationally intensive data-processing algorithms. As an example, consider the advanced process control. The real-time control of a valve would be operating at the temporal scale of seconds, and the decision making for the blending of end products at the top management level could be in a quarterly basis. The second reason is owing to the latest and fast development of microprocessor technology and parallel computing power, which greatly expedite the computation speed, and many off-line algorithms can now be executed in an on-line mode.

In fact, we anticipate this as a new trend in near future

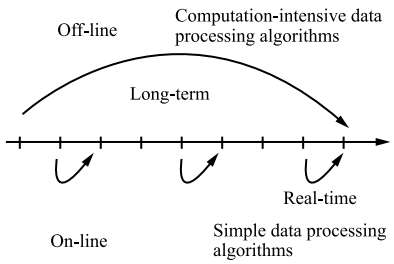

Fig. 8 An example with temporally multi-scale data processing

so that more sophisticated data-processing methods can be easily incorporated in real time. As a consequence, the boundary between off-line and on-line is getting blur, and more off-line data sets can be used fully or partially to support or enhance on-line tasks.

2 The complementary relationship between data and models

It is interesting to note that data-driven system approaches can also be model based or model driven. In fact, a physical model developed by the first principle can be viewed as an indirect data-driven model because many physical principles, such as Newton′s mechanics and Kirchhoff′s circuit laws, were invented and verified through experimental observations that were data based. The main differences between data and physical models lie in that the process information provided might be incomplete by the former but complete by the latter: the former need to go mining and learning to extract useful information contained in data, whereas the latter provides the useful information completely as the direct outcome. Unfortunately, physical models are difficult or impossible to build up for large-scale complex industrial processes. In addition, by aggregating hundreds and thousands of small-scale physical models, we still face difficulty to understand and handle the underlying characteristics of the large-scale industrial process. We need to seek and adopt a creative middle way solution. A possible alternative is to develop data-driven models that are made to function in a complementary manner to physical models. In industries, this middle way solution has been widely used as a part of their endeavors, some representative examples are rule base, lookup table, and graphics, which are extracted from off-line data.

Data-driven models can be classified into direct and indirect ones. The direct data-driven models aim at the input-output mapping, and off-line data are used. For instance, a neural network model can be tuned to fit the observed input-output data set. The neural network model, once established, can handle on-line data directly. Similar models include well structured filters, hidden Markov models, and other function approximation models such as Wavelet model. The indirect data-driven models aim at feature or attribute extraction from the observed data set. The feature extraction can be carried out using either data-mining or unsupervised machine learning methods such as K-means clustering, or using supervised learning with domain knowledge. When handling on-line data, some pre-filtering would be required to facilitate pattern recognition or template matching in feature or eigenspace. Fuzzy relation models, expert systems, probability density function (PDF) based models can be viewed as indirect data-driven models. An example of data-driven models[50] is shown in Fig. 9.

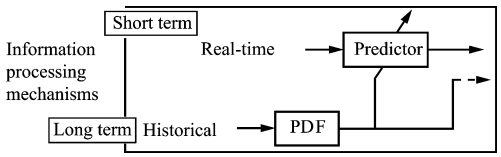

Fig. 9 An indirect data-driven model for trip forecasting (The PDF model is derived from historical trip data, which plays the role as a nominal model for long term forecasting; the predictor rectifies the outcome of the PDF model using on-line data.)

The disadvantage of data-driven modeling, which is obvious, links to the limits of on-line and off-line data, such as the incompleteness of data due to loss, uncertainty, bias, etc. An alternative is to combine data-driven models with physical models, subsequently form a hybrid model. The fusion of two classes of models can be carried out in different ways. First, for a process of interests, whenever a physical model is available, there is no need to develop a data-driven model. Data-driven models are required only when physical modeling is infeasible or too costly to develop. In this way, the data-driven model can be viewed as a subspace model to cover the missing parts of a physical model.

Second, when a large-scale industrial production process consists of many subsystems, it would be easier to model individual subsystems and a physical model can provide useful input-output relationship for a single component. However, we may lose insight on the interactions among multiple loops and the global behavior of a large-scale system, even though we could have the complete physical model for each of the components. In such circumstances, a data-driven model could be used to capture the predominant variables or principal components, decisive factors, and relations. Model reduction and characteristic model[51] are two representative approaches. The two key points for this kind of modeling approaches are the search for suitable model candidates and the selection of criteria for model validation. Domain knowledge plays a critical role, and therefore, niche areas are preferred. It remains an open issue on how to evolve into a universal modeling method.

Third, when an industrial production process is in a hierarchical structure, it is easier to model lower level components that consist of well defined physical variables and well known physical process models, for instance, at the level of actuation control, sensing, or component level fault identification and isolation. However, it may be difficult or impossible for us to find a meaningful global physical model at higher level, which actually involves non-physical factors such as human activities, cost and profit assessment, and soft constraints, for instance, at the level of scheduling with task assignment and optimization or at the level of planning with task generation and decision making. In such circumstances, data-driven models will be again useful and the vast amounts of off-line data can contribute to the establishment of such data-driven models.

Let us consider a large-scale industrial system, such as an advanced process control system, and see what kinds of data-driven models can be applied to address the four issues associated with control, decision making, scheduling, and fault diagnosis. A concept diagram of an advanced process control system is shown in Fig. 10, in which we omit the information flow from bottom upwards for simplicity.

First, look into the control issue using data-driven system approaches. The traditional controller design based on the physical model is shown in Fig. 11. Although the controller design can be carried out off-line, its effectiveness relies on the accuracy of the physical model, which is usually difficult to guarantee due to many factors such as high dimensions, nonlinearities, data loss, sampling limit, and randomness. A possible solution to this problem, as shown in Fig. 12, is to delink the physical model from the design model that is used for controller design. Since the ultimate objective of a controller is to meet control performance requirements instead of seeking modeling accuracy, we can use a simpler design model that is a lower order and reduced dimensional model in comparison with the original physical model, so long as the characteristic responses of the design model and physical model are consistent. In fact, this idea had been used in the transient response method for PID autotuing proposed by Ziegler-Nichols more than six decades ago[21], where the design model for PID setting is a simple first-order plus dead-time model regardless of the original physical process. By incorporating powerful data-mining and machine-learning methods developed in the past few decades, we can apply this method to much more complicated industrial systems, though there is still a long way to go. A more generic control system candidate is illustrated in Fig. 13.

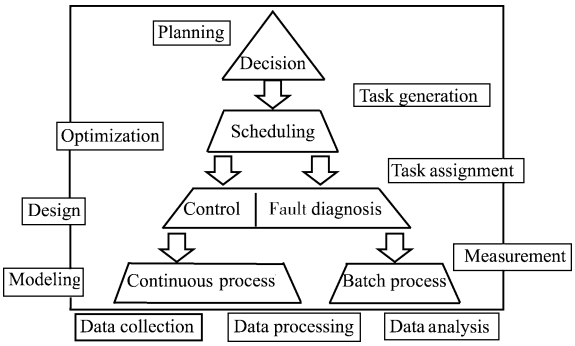

Fig. 10 A hierarchical structure of an advanced process control system (There are four levels: decision making, scheduling, control and fault diagnosis, continuous process and batch process. On the left, four objectives associated with each level are the planning, optimization, design, and process modeling, respectively. On the right, from top-down the three jobs in the four levels are task generation, task assignment, and measurement, respectively. At each level, data are collected when the system is in operation. Thus, the vast amounts of off-line data are available for processing, analyzing, and information acquisition.)

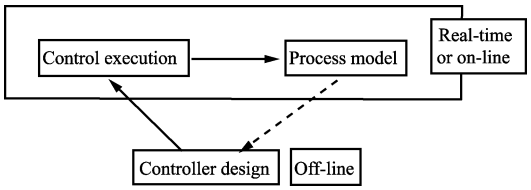

Fig. 11 A traditional controller design method in which the design model is the same as the physical process model

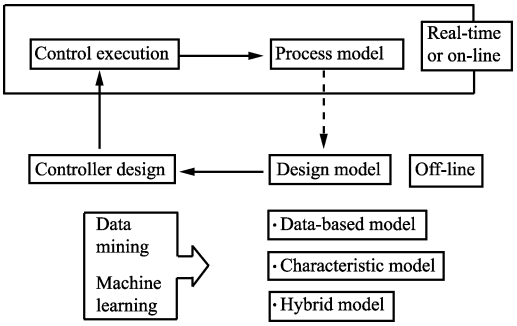

Fig. 12 An alternative controller design method in which the design model can be a lower order approximation of the actual physical process model

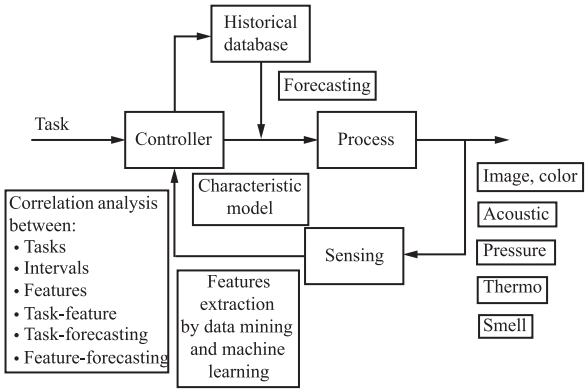

Fig. 13 A prototype of data-driven control systems that can sense and handle heterogeneous data, conduct feature analysis and extraction, perform correlation analysis, characteristic pattern matching, and forecasting

Second, look into the fault diagnosis issue using data-driven system approaches. A data-driven system approach for fault diagnosis and fault tolerance control is shown in Fig. 14. When multiple jobs are assigned to a number of manufacturing units, a hidden faulty unit would degrade the product quality. Through continuous monitoring and analysis, the change of product quality in terms of the PDF distribution can be detected. A non-Gaussian type distribution implies the presence of one or a few hidden faults. After the fault detection, fault types can be identified and faulty components can be isolated by means of the input-output mapping of each manufacturing unit. Based on the results of fault diagnosis, the fault tolerance controller will reassign jobs among normal units. It may happen that faults are not from a manufacturing unit but from some incorrect reference setting, and the direct retuning of setpoints will solve the problem. Here, fault tolerance control can be viewed as a part or scheduling tasks.

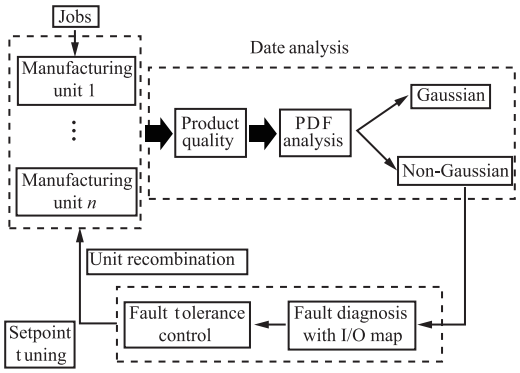

Fig. 14 A data-driven system approach for fault diagnosis and fault tolerance control

Next, look into the scheduling issue using data-driven system approaches. The block diagram of a suggested data-driven approach is shown in Fig. 15. Analogous to preceding case, the states of manufacturing units will be clustered in terms of feature extraction and analysis. Next, off-line learning methods, for example, the reinforcement learning, are employed to generate an input-output mapping or a decision machine, which can be either dynamic or stationary, of manufacturing units with job assignments as inputs and operation schedules as outputs. With the so-generated decision machine, real-time scheduling or rescheduling is practically implementable with respect to newly assigned jobs or unexpected variations.

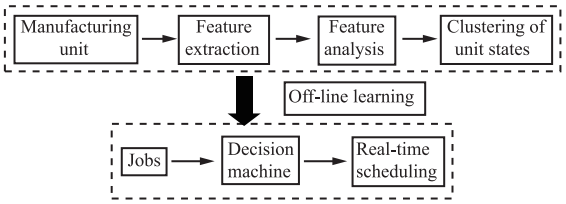

Fig. 15 A data-driven system approach for dynamic or real-time scheduling in presence of uncertain events or disturbances

Fig. 16 shows another possible schematic of data-driven system approaches for model-based monitoring and quality prediction. The off-line data is used to model the correlation among principal variables; accordingly the production process can be segmented into a number of intervals each with a state-space model. By virtue of each interval model, model-based on-line monitoring and prediction are performed.

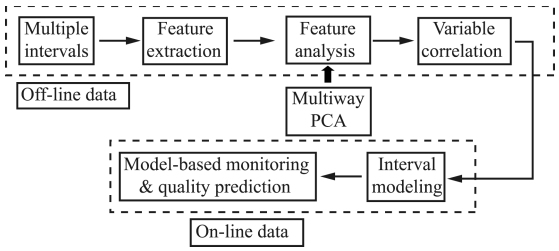

Fig. 16 A data-driven system approach for production process monitoring on safety, reliability, stable performance, and quality prediction, where the production process could be subject to uncertainties and disturbances

3 The perspectives of data-driven system approaches

Data-driven control, data-driven fault diagnosis, data-driven scheduling, and data-driven decision making present not only a new avenue but also new challenges both in theory and applications. In this section, we will explore the possibility to establish a theoretical framework for data-driven system methods. Three possible directions for explorations are extension, transfer and lifting, and innovation.

3.1 Extension

In the four key fields of interests, many theories have been proposed and developed. For instance, Fig. 5 shows nine main system control theories, each has attracted many dedicated researchers working on it. As a result, hundreds and thousands of articles and reports have been published on theory and applications. The research direction on data-driven system approaches can be chosen along the line of extension from existing system approaches.

Let us focus on system control. From Fig. 5, the nine system control approaches could be extended to data-driven PID, data-driven variable structure control, data-driven iterative learning control, and much more. A question arising is what kinds of attributes must be held by a data-driven control method. From discussions in preceding sections,

we can first conclude that a data-driven control method should be able to maximize the usage of historical or off-line data during selecting and designing controller structures and tuning controller parameters or be able to even directly incorporate the off-line data in the on-line control loops. Second, a data-driven model derived from off-line or on-line data could be another attribute of a data-driven system controller, where the process physical model is either not available or incomplete. Giving two examples, the data-driven model could be as simple as a gradient that updates with newly arrived on-line data, or an approximation model linearized around an operating point using historical data for the nominal part and on-line data for updating and fine tuning. The main open issue concerned with extension is the validity analysis of the extended data-driven system approaches, such as the stability and robustness analysis in data-driven system control. New analysis tools and system theories dedicated to the characteristics of data and data-driven models are necessary to support the new trend on data-driven systems. When data are with a PDF or a distributed reliability, we may also consider an extension principle analogous to fuzzy set theory.

3.2 Transfer and lifting

Computational intelligence, data-mining, machine learning, and many other application oriented research areas provide a rich source on data processing methods and algorithms for the four key research fields. Also, many methods, though originally developed for off-line data processing, can now be applied to on-line scenarios owing to the exploitation of computation-efficient algorithms and the evolution of microprocessor technology. Fig. 17 gives a list of seven such application oriented research areas: image processing, acoustic signal processing, time series forecasting, biology-inspired neural computing, bio-informatics, biomimetics, and biometrics, whose methods and techniques could be applied to solve problems in the four key fields of interests. As an example, Fig. 17 shows a list of subareas under biometrics: modeling and recognition of face, fingerprint, hand geometry, hand veins, iris, retinal scan, signature, voice, gait, ear canal, etc. In each subarea of biometrics, many effective feature extraction, identification, detection, isolation, classification, and recognition methods and techniques in the spatial, temporal, frequency, and time-frequency domains have been developed and even commercialized. What we need to do is to understand, digest, modify, and lift these methods and techniques for the purpose of control, decision making, scheduling, and fault diagnosis.

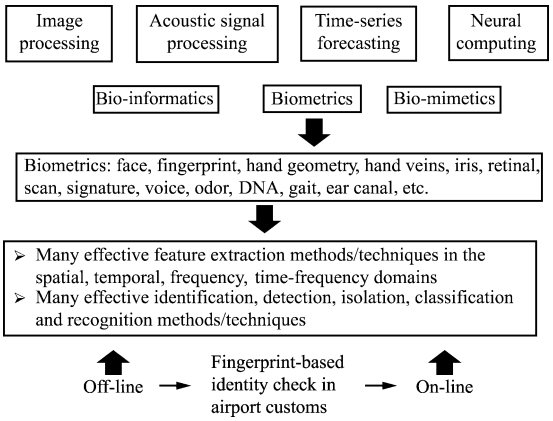

Fig. 17 Seven research areas that could be correlated to the four key fields of interests

Meanwhile, we should lift these ready-made approaches to fit dynamic processes in terms of the similarities in task characteristics, for example, lifting various forecasting methods for model-based prediction, dimension reduction methods for model reduction, and feature analysis methods for feature-based decision and control. As shown in Fig. 17, the fingerprint, originally developed for off-line processing, now is implemented in real time.

3.3 Innovation

Again let us focus on control. The methods and techniques extended from existing system control approaches would not exceed the existing scope of theory. Methods and techniques by transfer and lifting would be application specific or domain-knowledge dependent. Our ultimate goal is to explore and develop a new theoretical framework that can support universal data-driven system control approaches. A number of dedicated data-driven control methods, such as PID, iterative feedback tuning, iterative learning control, virtual reference feedback tuning, and model-free adaptive control are results of works and efforts made along this direction and are also introduced in this special issue[53]. Here, we will briefly touch the issue from another angle: can we define and develop fundamentals for data-centered system control in concept, theory, and designs. Considering the differences between physical-model based and data-driven approaches, it might be necessary to redefine system and control properties, such as shown in Fig. 18. Taking the controllability as an example, the practical controllability could refer to the controllability index of a data-driven model with the available off-line data at the design phase, or with the historical data up to certain time instance. Relative controllability could refer to the weakness of the controllability index quantified with a numerical value based on the available data set. Conditional controllability could be classified according to off-line or on-line data under certain conditions such as being pre-filtered with a given cutoff frequency. Ensemble controllability could refer to the controllability index under ensemble average. PDF-based controllability could be defined statistically according to the PDF of the available data set. Progressive controllability could be a dynamic index that updates with the newly arrival of on-line data. Likewise, we can apply the same or analogous attributes to other system properties, seek and provide more attributes that can better specify the data-driven system approaches from a unique angle.

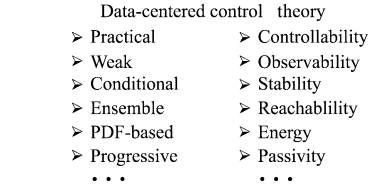

Fig. 18 Some possible redefinitions of data-driven system control properties (The left column represents possible data-driven attributes, and the right column represents control system properties.)

4 Conclusion

In this paper, we briefly discussed data-driven system problems and methods relevant to the four key fields on control, decision making, scheduling, and fault diagnosis. We first looked at the several data-processing methods and technology widely used in other research areas, which could

be applied to solve our problems. Next, we discussed the relationship between the off-line and on-line data, as well as the expected characteristics. Then, the data-driven and model driven system approaches were scanned with possible applications to the four key fields of interests. Finally, we explored the possibility to develop a new system theoretical framework.

In this paper, we do not intend to decide a new direction or a new avenue for the four key fields of interests, but we provide some primary concepts, suggest some possibilities that might be worth to explore, and recommend some other areas that could contribute to our problem solving.

Acknowledgement

Authors would like to thank NSFC organizers and participants who shared their ideas and works with us during the NSFC workshop on data-based control, decision making, scheduling, and fault diagnosis. In particular, authors would like to thank Chai Tian-You, Sun You-Xian, Wang Hong, Yan Hong-Sheng, and Gao Fu-Rong for discussing the concept on design model shown in Fig. 12, the concept on temporal multi-scale shown in Fig. 8, the concept on fault diagnosis shown in Fig. 14, the concept on dynamic scheduling shown in Fig. 15, and the concept on interval model shown in Fig. 16, respectively.

References

1 Hand D J, Mannila H, Smyth P. *Principles of Data Mining*. Cambridge: The MIT Press, 2000

2 Kononenko I, Kukar M. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Chichester: Horwood Publishing, 2007

3 Han J W, Kamber M. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann, 2006

4 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006

5 Holmes D E, Jain L C. *Innovations in Machine Learning: Theory and Applications*. Berlin: Springer, 2006

6 Jin Y C. *Multi-objective Machine Learning*. Berlin: Springer, 2006

7 Gurney K. *An Introduction to Neural Networks*. London: Routledge, 1997

8 Zadeh L A. *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems*. New York: World Scientific Press, 1996

9 De Jong K A. *Evolutionary Computation: a Unified Approach*. Cambridge: The MIT Press, 2006

10 Tan K K, Wang Q G, Hang C C. *Advances in PID Control*. London: Springer, 1999

11 Anderson B D O, Moore J B. *Optimal Control: Linear Quadratic Methods*. London: Prentice-Hall, 1989

12 Narendra K S, Annaswamy A M. *Stable Adaptive Systems*. London: Prentice-Hall, 1989

13 Sabanovic A, Fridman L, Spurgeon S K. *Variable Structure System: from Principles to Implementation*. New York: The Institution of Engineering and Technology, 2004

14 Jantzen J. *Foundations of Fuzzy Control*. New York: Wiley, 2007

15 Ge S S, Hang C C, Lee T H, Zhang T. *Stable Adaptive Neural Network Control*. Boston: Kluwer Academic, 2002

16 Zhou K M. *Robust Control of Uncertain Systems and H_∞ Optimization*. Ann Arbor: University Microfilms International, 1990

17 Findeisen R. *Assessment and Future Directions of Nonlinear Model Predictive Control*. New York: Springer, 2007

18 Xu J X, Tan Y. *Linear and Nonlinear Iterative Learning Control*. Berlin: Springer, 2003

19 Lee J M, Lee J H. Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes. *Automatica*, 2005, **41**(7): 1281−1288

20 Song Z, Kusiak A. Constraint-based control of boiler efficiency: a data-mining approach. *IEEE Transactions on Industrial Informatics*, 2007, **3**(1): 73−83

21 Wang X, Huang B, Chen T W. Multirate minimum variance control design and performance assessment: a data-driven subspace approach. *IEEE Transactions on Control Systems Technology*, 2007, **15**(1): 65−74

22 Sridhar P, Madni A M, Jamshidi M. Multi-criteria decision making in sensor networks. *IEEE Instrumentation and Measurement Magazine*, 2008, **11**(1): 24−29

23 Han J W, Lakshmanan L V S, Ng R T. Constraint-based multidimensional data mining. *Computer*, 1999, **32**(8): 46−50

24 Chaudhuri S, Dayal U, Ganti V. Database technology for decision support systems. *Computer*, 2001, **34**(12): 48−55

25 Togami M, Abe N, Kitahashi T, Ogawa H. On the application of a machine learning technique to fault diagnosis of power distribution lines. *IEEE Transactions on Power Delivery*, 1995, **10**(4): 1927−1936

26 Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: services, tools, and applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2004, **34**(6): 2451−2465

27 Murphey Y L, Crossman J A, Chen Z H, Cardillo J. Automotive fault diagnosis — part II: a distributed agent diagnostic system. *IEEE Transactions on Vehicular Technology*, 2003, **52**(4): 1076−1098

28 Yan L P, Liu B S, Zhou D H. An asynchronous multirate multisensor information fusion algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, 2007, **43**(3): 1135−1146

29 Polikar R. Essemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 2006, **6**(3): 21−45

30 Kusiak A, Kern J A, Kernstine K H, Tseng B T L. Autonomous decision-making: a data mining approach. *IEEE Transactions on Information Technology in Biomedicine*, 2000, **4**(4): 274−284

31 Chiu C C, Hsu K H, Hsu C I, Lee P C, Chiou W K, Liu T H. Mining three-dimensional anthropometric body surface scanning data for hypertension detection. *IEEE Transactions on Information Technology in Biomedicine*, 2007, **11**(3): 264−273

32 Liu J, You J. Smart shopper: an agent-based web-mining approach to internet shopping. *IEEE Transactions on Fuzzy Systems*, 2003, **11**(2): 226−237

33 Povinelli R J, Bangura J F, Demerdash N A O, Brown R H. Diagnostics of bar and end-ring connector breakage faults in polyphase induction motors through a novel dual track of time-series data mining and time-stepping coupled FE-state space modeling. *IEEE Transactions on Energy Conversion*, 2002, **17**(1): 39−46

34 Bangura J F, Povinelli R J, Demerdash N A O, Brown R H. Diagnostics of eccentricities and bar/end-ring connector breakages in polyphase induction motors through a combination of time-series data mining and time-stepping coupled FE-state-space techniques. *IEEE Transactions on Industry Applications*, 2003, **39**(4): 1005−1013

35 Wang Y Q, Zhou D H, Gao F R. Iterative learning model predictive control for multi-phase batch processes. *Journal of Process Control*, 2008, **18**(6): 543−557

36 Backus P, Janakiram M, Mowzoon S, Runger G C, Bhargava A. Factory cycle-time prediction with a data-mining approach. *IEEE Transactions on Semiconductor Manufacturing*, 2006, **19**(2): 252−258

37 Xu J X, Tan Y. Nonlinear adaptive wavelet control using constructive wavelet networks. *IEEE Transactions on Neural Networks*, 2007, **18**(1): 115−127

38 Liu D, Javaherian H, Kovalenko O, Huang T. Adaptive critic learning techniques for engine torque and air-fuel ratio control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 988−993

39 Wang A P, Afshar P, Wang H. Complex stochastic systems modelling and control via iterative machine learning. *Neurocomputing*, 2008, **71**(13-15): 2685−2692

40 Huynh V N, Nakamori Y, Ho T B, Murai T. Multiple attribute decision making under uncertainty: the evidential reasoning approach revisited. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 2006, **36**(4): 804−822

41 Qiang Y, Jie Y, Ling C, Rong P. Extracting actionable knowledge from decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 2007, **19**(1): 43−56

42 Wang G, Gong W R, DeRenzi B, Kastner R. Ant colony optimizations for resource- and timing-constrained operation scheduling. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2007, **26**(6): 1010−1029

43 Yang S, Wang D W. Constraint satisfaction adaptive neural network and heuristics combined approaches for generalized job-shop scheduling. *IEEE Transactions on Neural Networks*, 2000, **11**(2): 474−486

44 Togami M, Abe N, Kitahashi T, Ogawa H. On the application of a machine learning technique to fault diagnosis of power distribution lines. *IEEE Transactions on Power Delivery*, 1995, **10**(4): 1927−1936

45 Khomfoi S, Tolbert L M. Fault diagnosis and reconfiguration for multilevel inverter drive using AI-based techniques. *IEEE Transactions on Industrial Electronics*, 2007, **54**(6): 2954−2968

46 Zhu Y L, Limin H, Lu J L. Bayesian networks-based approach for power systems fault diagnosis. *IEEE Transactions on Power Delivery*, 2006, **21**(2): 634−639

47 Chia H W K, Tan C L, Sung S Y. Enhancing knowledge discovery via association-based evolution of neural logic networks. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18**(7): 889−901

48 Tsipouras M G, Exarchos T P, Fotiadis D I, Kotsia A P, Vakalis K V, Naka K K. Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling. *IEEE Transactions on Information Technology in Biomedicine*, 2008, **12**(4): 447−458

49 Chien C F, Huang Y C, Hu C H. A hybrid approach of data mining and genetic algorithms for rehabilitation scheduling. *International Journal of Manufacturing Technology and Management*, 2009, **16**(1-2): 76−100

50 Lin W H. A Gaussian maximum likelihood formulation for short-term forecasting of traffic flow. In: Proceedings of IEEE Intelligent Transportations Systems. Oakland, USA: IEEE, 2001. 150−155

51 Wu H X, Liu Y W, Liu Z H, Xie Y C. Characteristic modeling and the control of flexible structure. *Science in China (Series F): Information Sciences*, 2001, **44**(4): 278−291

52 Ziegler J G, Nichols N B. Optimum settings for automatic controllers. *Journal of Dynamic Systems, Measurement, and Control*, 1993, **115**(2): 220−222

53 Hou Zhong-Sheng, Xu Jian-Xin. On data-driven control theory: the state of the art and perspective. *Acta Automatica Sinica*, 2009, **35**(6): 650−667

XU Jian-Xin Received his bachelor degree from Zhejiang University in 1982. He then attended the University of Tokyo, Japan, where he received his master and Ph. D. degrees in 1986 and 1989, respectively. All his degrees are in electrical engineering. He worked for one year in the Hitachi Research Laboratory, Japan; for more than one year in Ohio State University, USA as a visiting scholar; and for 6 months in Yale University, USA as a visiting research fellow. In 1991, he joined the National University of Singapore, Singapore, and is currently a professor in the Department of Electrical Engineering. His research interest covers learning control, variable structure control, fuzzy logic control, discontinuous signal processing, and applications to motion control and process control. Corresponding author of this paper. E-mail: elexujx@nus.edu.sg

HOU Zhong-Sheng Received his bachelor and master degrees in applied mathematics from Jilin University of Technology in 1983 and 1988, respectively, and his Ph. D. degree in control theory from Northeastern University in 1994. From 1988 to 1992, he was a lecturer with the Department of Applied Mathematics, Shenyang Polytechnic University. He was a postdoctoral fellow with Harbin Institute of Technology from 1995 to 1997, and a visiting scholar with Yale University, New Haven, CT, from 2002 to 2003. In 1997, he joined Beijing Jiaotong University and is currently a professor with the Department of Automatic Control. His research interest covers the model free adaptive control, learning control, and intelligent transportation systems.
E-mail: zhshhou@bjtu.edu.cn