Computer Physics Communications 234 (2019) 159–166

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Time-domain global similarity method for automatic data cleaning for multi-channel measurement systems in magnetic confinement fusion devices

Ting Lan [a], Jian Liu [a,*], Hong Qin [a,b], Lin Li Xu [c]

[a] School of Nuclear Science and Technology and Department of Modern Physics, University of Science and Technology of China, Hefei, Anhui 230026, China
[b] Plasma Physics Laboratory, Princeton University, Princeton, NJ 08543, USA
[c] School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China

ARTICLE INFO

Article history:
Received 8 October 2017
Received in revised form 7 February 2018
Accepted 21 July 2018
Available online 25 July 2018

Keywords:
Machine learning technique
Automatic data cleaning
Time-domain global similarity
MUlti-channel measurement system
Magnetic confinement fusion device

ABSTRACT

To guarantee the availability and reliability of data source in Magnetic Confinement Fusion (MCF) devices, incorrect diagnostic data, which cannot reflect real physical properties of measured objects, should be sorted out before further analysis and study. Traditional data sorting cannot meet the growing demand of MCF research because of the low-efficiency, time-delay, and lack of objective criteria. In this paper, a Time-Domain Global Similarity (TDGS) method based on machine learning technologies is proposed for the automatic data cleaning of MCF devices. The aim of traditional data sorting is to classify original diagnostic data sequences. The lengths and evolution properties of the data sequences vary shot by shot. Hence the classification criteria are affected by many discharge parameters and are different in various discharges. The focus of the TDGS method is turned to the physical similarity between data sequences from different channels, which are more independent of discharge parameters. The complexity arisen from real discharge parameters during data cleaning is avoided in the TDGS method by transforming the general data sorting problem into a binary classification problem about the physical similarity between data sequences. As a demonstration of its application to multi-channel measurement systems, the TDGS method is applied to the EAST POlarimeter–INTerferometer (POINT) system. The optimal performance of the method evaluated by 24-fold cross-validation has reached 0.9871 ± 0.0385.

1. Introduction

With the development of Magnetic Confinement Fusion (MCF) science and diagnostic techniques, massive diagnostic data are increasingly generated. Original diagnostic data could be unreliable due to various interference sources and complex measuring conditions in MCF devices, such as mechanical vibration, electromagnetic interference, signal saturation, and hardware failures. To guarantee the availability and reliability of data source, incorrect diagnostic data, dubbed dirty data, which cannot reflect real physical properties of measured objects, should be sorted out before further analysis and study. The identification of incorrect data can be regarded as a typical classification problem, i.e., how to properly divide the original dataset into two groups, correct data and incorrect ones. Since experimental setups and discharge processes are diverse, measured quantities from different shots, diagnostic systems, and devices evolve in totally different ways. Incorrect diagnostic results also vary due to their uncertain causes. Therefore, it is difficult to define general and clear criteria for data cleaning. Traditionally, dirty data are searched and removed manually with the assistant of computer programs, mainly according to some simple rules, common experiences, and sometimes personal intuitions. These data cleaning programs and rules only apply to certain specific data and usually perform poorly. Explosively increasing fusion data cannot be satisfactorily cleaned in time. Real-time processing and feed-back control require much faster data cleaning methods, which can remove dirty data in a short time, e.g., a few milliseconds. On the other hand, subjective factors in manual data cleaning processes lead to inconsistent results. To meet the demand of fusion energy research, the speed, efficiency, and accuracy of fusion data cleaning should be improved imperatively. Automatic data cleaning methods based on machine learning is a strong candidate for breaking through the bottleneck of massive data application in fusion research.

In recent years, as computing ability and storage capacity grow rapidly, Artificial Intelligence (AI) and machine learning have been widely applied to a variety of scientific research fields, such as image processing, biology, and astronomy [1–3], showing great advantages of extracting new patterns and principles from

[*] Corresponding author.
E-mail address: jliuphy@ustc.edu.cn (J. Liu).

https://doi.org/10.1016/j.cpc.2018.07.014
0010-4655/© 2018 Published by Elsevier B.V.

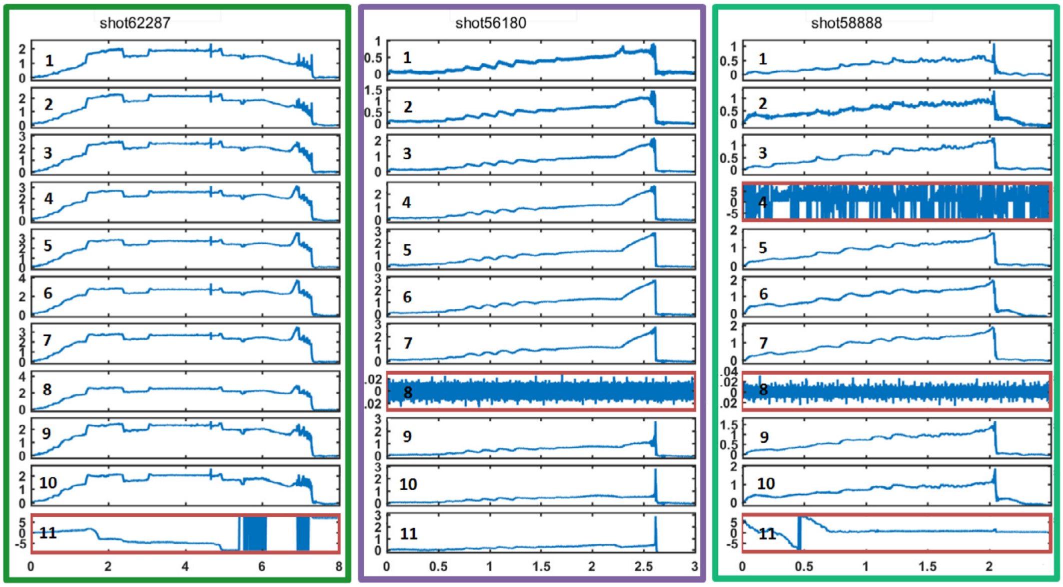

Fig. 1. Original density sequences from different channels of the EAST POlarimeter–INTerferometer (POINT) system are plotted for three typical discharges {shot 62 287, 56 180, 58 888}. The channel id of the sequences is labeled. The channel id of incorrect sequences is {11} for shot 62 287, {8} for shot 56 180, and {4, 8, 11} for shot 58 888, respectively. These incorrect sequences are marked with red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

complicated big dataset. In MCF research, machine learning has been applied in disruption prediction [4–9], plasma fluctuations extraction [10], data retrieval [11], L–H transition time estimation [12], charge exchange spectra analysis [13], neoclassical transport database construction [14], turbulent transport construction [15], electron temperature profile reconstruction [16], and energy confinement scaling [17]. These pioneering works have pushed the fusion energy research forward in many aspects, respectively. Wider, larger-scale, and systematic application of machine learning in fusion science will trigger radical changes. Effective data cleaning also becomes the essential prerequisite for the application of AI, data mining, and big data techniques in fusion research. Machine learning in turn offers powerful tools for diagnostic data cleaning. Precise data cleaning can be achieved by using objective classification model trained by original data using supervised machine learning methods. The application speed of well-trained model will be easily optimized to meet the requirements of real-time feedback control. With the support of supercomputer, massive fusion data can be processed effectively to relieve the data processing pressures of researchers. The robustness and universality of classification models lays a foundation for the large-scale applications of machine learning in fusion science.

In this paper, a new data cleaning method based on the Time-Domain Global Similarity (TDGS) among data sequences defined by typical machine learning technologies is proposed. The general-purposed TDGS method can be used to automatically sort dirty diagnostic data from MUlti-channel Measurement (MUM) systems in MCF experiments. Most diagnostic systems of MCF devices are MUM systems, which measure the time evolution of plasma parameters from different locations or directions with multiple independent measuring channels, such as common interferometer systems [18], polarimeter systems [19–23], and electron cyclotron emission imaging systems [24]. Time sequences of diagnostic data from different channels of the MUM system reflect related yet distinct aspects of the same observed object. Therefore these diagnostic data are physically associated. We define this relation as physical similarity. The physical similarity just exists between correct data sequences from different channels of the MUM system. The dirty data, which are caused by a variety of interference sources, are physically dissimilar from correct data sequences or each other. To overcome the difficulty of direct classification, the TDGS method sorts the dirty data by classifying the physical similarity between diagnostic data sequences from different channels under the same discharge. The aim of traditional data sorting is to classify original diagnostic data sequences. The lengths and evolution properties of the data sequences vary shot by shot. Hence the classification criteria are affected by many discharge parameters and are different in various discharges. The goal of the TDGS algorithm is to classify the physical similarity between data sequences from different channels. This physical similarity is more independent of discharge parameters. Then the complexity arisen from real discharge parameters during data cleaning is avoided in the TDGS method by transforming the general data sorting problem into a binary classification problem about the physical similarity between data sequences.

In the TDGS method, the sample set is generated by combining two original diagnostic data sequences from two different channels of a MUM system in the same discharge. By combining two data sequences from different channels of an N-channel MUM system as one sample, C_N^2 samples can be generated for one discharge, and $P * C_N^2$ samples can be generated for P discharges. Each sample is tagged by several indices which indicates the corresponding physical similarity between two sequences. These indices span a feature space, in which these samples can be classified into two groups, physically similar samples and physically dissimilar ones. A physically similar sample is constituted by two correct data sequences. If a sample is classified to be physically dissimilar, its constituents contain at least one dirty data sequence. According to this rule, the dirty diagnostic data can be properly identified by physical similarity. In many MUM systems of MCF devices, the physical similarity between diagnostic data exists in time domain rather than frequency domain. As shown in Fig. 1, the incorrect data are dissimilar from each other, and from the correct data. In most cases, this dissimilarity has global characteristics, instead of local or small-scale features. The TDGS method employs different definitions of the distance between two time-series signals as tag indices of a sample, measuring this global time-domain similarity. To guarantee precise classification, different kinds of distance functions are adopted to map signals from the space of