Information Sciences 459 (2018) 149–167

Contents lists available at ScienceDirect

Information Sciences

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

INFORMATION SCIENCES

Informatics and Computer Science

Intelligent Systems

Applications

AN INTERNATIONAL JOURNAL

Including Special Section

Nature-Inspired Algorithms for Large Scale Global Optimization

Guest Editors

Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang

Available online at www.sciencedirect.com

ScienceDirect

Learning in the compressed data domain: Application to milk quality prediction

Dixon Vimalajeewa[a,*], Chamil Kulatunga[a], Donagh P. Berry[b]

[a] *Telecommunications Software and Systems Group, Arclabs Research and Innovation Centre, Waterford Institute of Technology, Carriganore, Waterford, Ireland*

[b] *Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy, Co. Cork, Ireland*

ARTICLE INFO

Article history:
Received 5 July 2017
Revised 1 May 2018
Accepted 2 May 2018

Keywords:
Compressed learning
MIRS
Principal component analysis
Wavelet transformation
Partial least squares regression
Fog computing

ABSTRACT

Smart dairy farming has become one of the most exciting and challenging area in cloud-based data analytics. Transfer of raw data from all farms to a central cloud is currently not feasible as applications are generating more data while internet connectivity is lacking in rural farms. As a solution, Fog computing has become a key factor to process data near the farm and derive farm insights by exchanging data between on-farm applications and transferring some data to the cloud. In this context, learning in the compressed data domain, where de-compression is not necessary, is highly desirable as it minimizes the energy used for communication/computation, reduces required memory/storage, and improves application latency. Mid-infrared spectroscopy (MIRS) is used globally to predict several milk quality parameters as well as deriving many animal-level phenotypes. Therefore, compressed learning on MIRS data is beneficial both in terms of data processing in the Fog, as well as storing large data sets in the cloud. In this paper, we used principal component analysis and wavelet transform as two techniques for compressed learning to convert MIRS data into a compressed data domain. The study derives near lossless compression parameters for both techniques to transform MIRS data without impacting the prediction accuracy for a selection of milk quality traits.

1. Introduction

Even though smart farming is advancing with the recent developments of Internet of Things (IoT), cloud-based computing, and deep learning, it has become one of the most challenging industrial sectors in big data analytics due to the limitations of ICT infrastructures [47]. However, according to the statistics from the Food and Agriculture Organization of the United Nations (FAO), smart farming will be a key contributor to sustainable intensification in agriculture to feed the 9.2 billion human population by 2050 [1]. There is also a growing interest in pasture-based smart dairy farming in the countries like New Zealand and Ireland, which tend to be in less direct competition with human edible protein and energy sources. Therefore, more harmonized research is needed to optimally utilize ICT infrastructures in precision dairy farming to minimize consumed storage space, communication and computations to facilitate contemporary analytics providing near real-time insights on-farm [37]. This is where the notion of effective data compression approaches are important.

[*] Corresponding author.
E-mail addresses: dvimalajeewa@tssg.org (D. Vimalajeewa), ckulatunga@tssg.org (C. Kulatunga), Donagh.Berry@teagasc.ie (D.P. Berry).

https://doi.org/10.1016/j.ins.2018.05.002

Most sensor-based technologies and IoT platforms are designed today to collate and store vast quantities of raw data readings from different sources in geographically distributed farms. Many computational facilities for data analytic applications such as MyAgCentral[1] are now seeking computational resources in cluster-based servers in large centralized data centres. At same the time, the Agricultural Information Management Standards of FAO (AIMS) has already started developing standards and maintaining interoperable trans-national databases for open agricultural data. Therefore farm data will be aggregated as big datasets and there is a requirement to store these data for long-term analytical purposes. This is beneficial since aggregation of data, which extracts a large number of descriptive features in temporally and spatially diverse domains, contributes to an improved learning accuracy. Therefore, compression of such data without a loss of accuracy is vital in terms of the *storage* requirement.

Dissemination of data in its raw format (i.e., in the measurement domain) into large cloud-based data centres is generally not feasible for most farms due to high energy consumption, time criticality of the applications, and the poor/costly rural internet connectivity. For example, if a disease detection system is centralized, it may slow down the farmers' response because of the necessity to transfer vast quantity of data readings to a remote cloud and wait for the outcome to return. Therefore, compression of data is also important in terms of the *communication* efficiency.

However, the key challenge today is whether the centralized storage and computational technologies (with communication networks) contributing to smart dairy farming will still not be sufficient to deliver the future demand without an advanced data analytic infrastructure closer to remote farm management systems. Therefore, a scalable computational infrastructure under constrained resources (proximate to the farm) is essential. In such a constrained infrastructure, compression is a key performance factor also for *computational* efficiency in addition to storage and communication.

Emergence of Fog Computing: With the increase in the amount of data generated from connected sensors, there is a demand to move processing capabilities closer to the data sources, which is in contrast to centralizing raw data in a large data centre. This phenomenon of distributing computations towards the data was first termed as data gravity by *Dave McCrory* in 2010 and is now being realized with new technologies such as Fog (i.e., edge) computing [3] and cloudlets [22]. Fog computing can enable datasets to be processed at the extreme edge of the internet. This computational infrastructure may collectively be formed by low computational proximate devices located near or within the farm. Therefore Fog computing will be a key enabler for many farm analytics to run using scalable in-memory data processing platforms like Spark,[2] Flink,[3] Storm[4] and H20[5] with in-memory databases like Ignite[6] and SAP HANA. Therefore raw-data compression near the data source is a desirable requirement for near future.

As a result, machine learning models, which have targeted highly-provisioned cloud infrastructures, must be re-designed for these resource-constrained infrastructures to minimize storage, communication and computational requirements. New distributed machine learning paradigms like compressed learning [4] and attribute-distributed learning [50] have significant potential to develop effective learning models [49] rather than centralizing all raw datasets from the farms. The main motivation of the present paper is to validate a compressed learning approach [4] for milk quality analysis based on Mid-infrared spectroscopy (MIRS) technology, which can effectively overcome those three challenges in Fog computing. With compressed learning, any machine learning algorithm can be used in a low-dimensional (i.e. latent) space without decompressing data, while optimizing the resource requirement as well as learning efficiency and accuracy of outcomes. Even though the compressed learning approach has been widely used in many fields for learning from complex data sources, such as high-resolution image and video processing and text analysis [26,31], its applicability is new to the MIRS based milk quality analysis.

MIRS is the most economical technology used for assessing milk quality. Therefore, MIRS spectra in predictive models are frequently used to develop farm decision-support tools for efficient milk data processing. For instance, the OptiMIR project has used MIRS of milk recordings in an innovative way to observe different characteristics of cows such as energy balance and early detection of diseases. Also, the routinely obtained MIRS of milk can be used for deriving novel models to quantify milk composition on both an animal basis and on bulk tank samples as well as derive milk related herd-level phenotypes [29]. In addition, variation in MIRS of milk can be used as an indicator in predicting animal characteristics such as the physiological state of an animal and its feed efficiency. The collaborative use of MIRS milk data from different farms can also improve the accuracy of the predictions. Therefore processing vast quantities of milk samples with Fog computing is highly desirable for MIRS milk quality analysis in the future smart dairy farming.

Conventional MIRS analysis [42,43] has been conducted based on co-located data processing by a single computational facility. As shown in Fig. 1, the raw data are directly collated into a repository, mostly by non-experts of data science, and later analysed by the domain-specific data science experts. This significantly increases the computation and power resource requirement on the cloud using raw MIRS data. In modern distributed processing infrastructures, data pre-processing such as de-noising and dimensionality reduction can be carried out closer to data sources. It would, in turn, improve three forms of resource efficiency of the system and reduce the input cost compared to the conventional approach. Water absorbance

[1] https://myagcentral.com/.
[2] http://apache.org.
[3] http://www.h2o.ai.
[4] https://www.keep.eu/keep/project-ext/21114/OptiMIR.
[5] http://succinct.cs.berkeley.edu.
[6] http://CRAN.R-project.org/package=wavelets.