International Journal of Approximate Reasoning 88 (2017) 1–22

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A three-way decisions model with probabilistic rough sets for stream computing ☆

Jianfeng Xu [a,b,c], Duoqian Miao [a,c,*], Yuanjian Zhang [a,c], Zhifei Zhang [a,d]

[a] Department of Computer Science and Technology, Tongji University, Shanghai 201804, P.R. China
[b] Software College, Nanchang University, Jiangxi 330077, P.R. China
[c] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, P.R. China
[d] Big Data and Network Security Research Center, Tongji University, Shanghai 200092, P.R. China

A R T I C L E I N F O

Article history:
Received 29 June 2016
Received in revised form 30 December 2016
Accepted 8 May 2017
Available online 11 May 2017

Keywords:
Three-way decisions
Probabilistic rough sets
Stream computing learning method
Knowledge updating

A B S T R A C T

Stream computing paradigm, with the characteristics of real-time arrival and departure, has been admitted as a major computing paradigm in big data. Relevant theories are flourishing recently with the surge development of stream computing platforms such as Storm, Kafka and Spark. Rough set theory is an effective tool to extract knowledge with imperfect information, however, related discussions on synchronous immigration and emigration of objects have not been investigated. In this paper, stream computing learning method is proposed on the basis of existing incremental learning studies. This method aims at solving challenges resulted from simultaneous addition and deletion of objects. Based on novel learning method, a stream computing algorithm called single-object stream-computing-based three-way decisions (SS3WD) is developed. In this algorithm, the probabilistic rough set model is applied to approximate the dynamic variation of concepts. Three-way regions can be determined without multiple scans of existing information granular. Extensive experiments not only demonstrate better efficiency and robustness of SS3WD in the presence of objects streaming variation, but also illustrate that stream computing learning method is an effective computing strategy for big data.

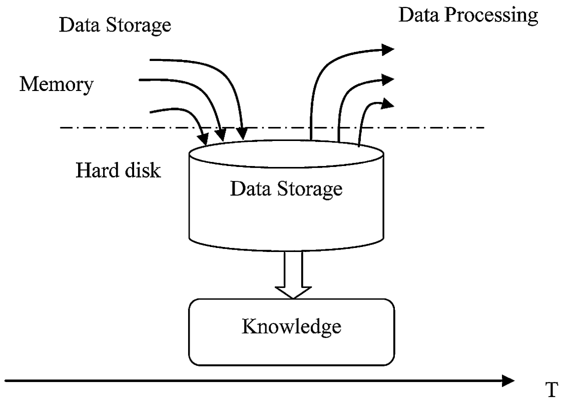
1. Introduction

The unprecedented popularity of novel information technology and application schema, such as cloud computing, Internet of things (IOT), and mobile interconnection, accumulate a large scale of data and promote the development of big data [2,7,39]. The essential characteristics of big data have been summarized by many scientists, and currently 5V model and 5R model [8] are widely accepted. Generally speaking, data with any properties or requirements mentioned in 5V and 5R model can be considered as big data. Fast arrival, for example, is admitted as one of the most remarkable challenges. On one hand, desirable result of up-to-date data cannot be achieved in a limited time because of high velocity (defined in 5V), on the other hand the value of complicated applications [13,15,33] will be diminished if hidden knowledge is not extracted real-time (defined in 5R). Specifically, there are two major reasons:

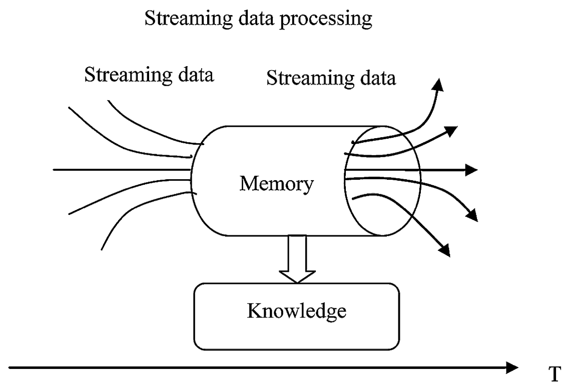
☆ This paper is part of the virtual special issue on tri-partition, edited by Davide Ciucci and Yiyu Yao.

* Corresponding author at: Department of Computer Science and Technology, Tongji University, Shanghai 201804, P.R. China.
E-mail address: dqmiao@tongji.edu.cn (D. Miao).

http://dx.doi.org/10.1016/j.ijar.2017.05.001
0888-613X/



(a) Batch computing paradigm



(b) Stream computing paradigm

Fig. 1. Typical computing paradigm for big data [37].

The explosive generation of data in a limited time is ubiquitous. For example, CERN's large hadron collider produce petabytes of data per second in the working status. Repeating such experiment is quite costly and thus one-pass scan of such application is very important.

In some other applications, however, data is poured from a tremendous number of interacting instances, despite the seemingly negligible contribution of each participant. Typical examples of this style include click stream, RFID data and GPS location information.

Obviously, these kind of data need to be processed in the manner of *stream computing*, i.e. computed the whole data segmentally and sequentially. To solve the dilemma of real-time and accurate, scholars have suggested a wealth of ideas and they can be categorized into two computing paradigms: *batch computing paradigm* and *stream computing paradigm*.

(1) Batch computing paradigm [46] stores and computes data in batches, whereas relations between batches are neglected. In most cases, both stages are handled in a highly centralized way. As depicted in Fig. 1(a), batch computing launches only when accumulated data is abundant.

(2) Stream computing paradigm [31], however, performs data storage and computing in memory and consider the relation of batches in the way of sliding window. As described in Fig. 1(b), no data exchange occurs between hard disk and memory.

Generally, any machine learning method can be customized into batch learning paradigm and stream computing paradigm. Incremental learning method can accelerate the speed of stream data. However, most of existing incremental learning method compute information addition and deletion separately, although both operations may be considered [14]. To implement stream learning paradigm more effectively, we term a new glossary called *stream computing learning method*.

Stream computing learning method is defined as a novel strategy whose operations towards objects immigrations and emigrations are conducted at the same time.

Obviously, stream computing learning method is more consistent to the connotation of stream computing paradigm as compared to classical incremental learning method. Currently, researches on the variation mechanism of data, fast real-time computing and approximate real-time computing are rather preliminary. It is undoubtedly that approximation instead of accurate answer is more likely to be achieved, therefore it is imperative to introduce new theory to facilitate the research of stream computing learning method.

Three-way decisions theory [43] is an important extension of rough set theory [34]. Decisions are determined if it is informative, otherwise will be deferred. It is a rather inclusive paradigm since the hidden structure used to support decision-making can be generated by any kinds of learning mechanism. Gradually, it has been recognized that the theory has incomparable advantages in solving complicated problem because of analogous cognitive mechanism shadowed in human [22,32,45]. Currently, the research direction of three-way decisions are mainly concentrated on the following aspects: 1) the basic theory of three-way decisions [9,11,16,17]; 2) three-way decision and rough sets theory [18,41,44,53,54]; and 3) clustering/classification based on three-way decisions [47–49].

The contributions of this paper are as follows. Firstly, it is the first time to systematically clarify the hierarchical structure of stream computing. From the coarsest to refinement, we have stream computing paradigm, stream computing learning method, and stream computing learning algorithms. Major differences against incremental learning lies in the level of stream computing learning method. While incremental learning method performs computations in the unit of information variation direction of dataset, i.e. either addition or deletion, stream computing learning method combines both immigrations and emigrations into an atomic operation unit. It is straightforward to see that the stream computing learning method is tightly coupled as compared to incremental learning method. Consequently, detailed stream computing learning algorithm is also different from incremental learning algorithm, making the knowledge updating more purposeful. Secondly, the use