Procedia Computer Science

Volume 53, 2015, Pages 9–18

2015 INNS Conference on Big Data

Evolving Classifier TEDAClass for Big Data

Dmitry Kangin*, Plamen Angelov*, **, Jose Antonio Iglesias***, Araceli Sanchis ***

* Data Science Group, Computing & Communications, Lancaster University, UK {d.kangin, p.angelov}@lancaster.ac.uk

** Chair of Excellence, Carlos III University of Madrid, Spain

*** Carlos III University of Madrid, Computer Science Department, Spain

Abstract

In the era of big data, huge amounts of data are generated and updated every day, and their processing and analysis is an important challenge today. In order to tackle this challenge, it is necessary to develop specific techniques which can process large volume of data within limited run times.

TEDA is a new systematic framework for data analytics, which is based on the *typicality* and *eccentricity* of the data. This framework is spatially-aware, non-frequentist and non-parametric. *TEDA* can be used for development of alternative machine learning methods, in this work, we will use it for classification (*TEDAClass*). Specifically, we present a *TEDAClass* based approach which can process huge amounts of data items using a novel parallelization technique. Using this parallelization, we make possible the scalability of *TEDAClass*. In that way, the proposed approach is particularly useful for various applications, as it opens the doors for high-performance big data processing, which could be particularly useful for healthcare, banking, scientific and many other purposes.

Keywords: Big Data, TEDA, AnYa, Evolving Systems for Big Data Analytics

1. Introduction

Huge amounts of data are generated every day in the modern society, mostly in a digital form. This makes the old approach to store all data items and for further processing and analysis impossible and gives rise to the term big data. Big data can be defined as a scale of dataset that goes beyond existing database management tool capabilities of data collection, storage, management, and analysis capabilities [1]. Although, the most common trait of big data is Volume, it is typically defined by more Vs such as Volume, Variety, Velocity, Veracity, Volatility, etc.

There are many different applications in which Big Data techniques are applicable: data mining, predictive analytics, geo-analysis, natural language processing and pattern recognition. Nowadays, since the computational power and reliability of the contemporary computers continue to grow, the data mining problems for big data is becoming widespread. These problems are shifting from the instrument for government, large corporations, banks to mass users. Big Data can be classified taking into account

E-mail addresses: d.kangin@lancaster.ac.uk (Dmitry Kangin), p.angelov@lancaster.ac.uk (Plamen Angelov), jiglesia@inf.uc3m.es (José Antonio Iglesias), masm@inf.uc3m.es (Araceli Sanchis).

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.274

the data type: (1) Structured (data are stored in fixed field), (2) Semi-structured (data are not stored in fixed field but the data includes metadata or schema), and (3) Unstructured (data are not stored in fixed field).

While processing structured and semi-structured data poses problems related primarily to the storage, retrieval and tagging, when unstructured data is concerned the primary problem is organising and making sense from it. Nowadays, unlike in the past century, vast majority of the data is unstructured. Simpler algorithms, such as retrieval, search and clustering can be approached using Map-Reduce \cite{SinghReddy14jobd} and through parallelisation can be scaled into a number of processing units (PU). More complex machine learning algorithms, however, such as classification, prediction, image processing etc., which are often iterative, are significantly more difficult to parallelise and scale up.

In this paper, we propose an algorithm for classification of Big Data based on the recently introduced TEDAClass (Typicality and Eccentricity based Data Analytics) approach [2]. TEDAClass itself is a neuro-fuzzy classifier which can be of zero or first order. The zero order TEDAClass has the class label as output. The first order version is using a mixture of (linear) regression models combined by a fuzzy weight proportional to their local density [3]. TEDAClass is based on the recently proposed alternative data analytics called TEDA [2] [4] [5].

It is important to stress that the proposed approach has been designed as a context-independent approach. It means that it is not restricted to any particular application. The proposed approach also does not have restrictive prior assumptions that are typical for alternative statistical, fuzzy rule-based and other approaches. This is due to being completely data-driven and based on the data density derived from data items.

This paper is organized as follows: In the next section, the background and related work to the proposed problem are discussed. In addition, the TEDA framework on which the proposed approach is based on is also outlined. Section 3 details the structure of the proposed parallelization approach (called $\text{TEDAClass}_{\text{BDp)}}$. Section 4 describes initially an intuitive illustrative example (IRIS classification data), and one larger realistic approach (ETL1 data set) presents the experimental settings and the obtained results. Finally, Section 5 makes the conclusions and outlines the future work.

2. Background and Related Work

Different scientific fields are becoming increasingly data-driven which also requires new approaches to be developed within the Computer Science to reflect this. For example, social computing [6] is becoming a discipline on its own; same is true for the bioinformatics [7], econometrics, astronomy is increasingly data-driven [8], etc. are examples of these fields. Big data require new type of computational approaches and techniques in order to be able to process efficiently large volumes of data within limited run times.

Various approaches where proposed specifically for Big Data recently, like those described in [9]. Perhaps the most popular approach is Hadoop and Map-reduce [10], but it is applicable to simpler problems such as information storage and retrieval, clustering, samples-centred approaches such as SVM etc. Most of the Big Data specific approaches are adaptations of previously existing and well known machine learning approaches, for example, k-means clustering [11], SVM [12], fuzzy and probabilistic clustering Fuzzy logic algorithms [13].

In the following subsections the two important aspects of the proposed approach are explained in more detail: i) the parallelization structure, and ii) the *TEDA* classifier.