Future Generation Computer Systems 35 (2014) 49–56

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A tensor-based distributed discovery of missing association rules on the cloud

Isam Elayyadi[a], Salima Benbernou[a], Mourad Ouziri[a], Muhammad Younas[b,*]

[a] Université Paris Descartes, Sorbonne Paris Cité, France

[b] Department of Computing and Communication Technologies, Oxford Brookes University, Oxford, UK

HIGHLIGHTS

- Distributed frequent itemsets.
- Data aggregation on the cloud.
- Repairing missing data.

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form
20 September 2013
Accepted 14 November 2013
Available online 25 November 2013

Keywords:
Distributed mining
Association rules
Cloud computing
Tensor

ABSTRACT

An increasing number of data applications such as monitoring weather data, data streaming, data web logs, and cloud data, are going online and are playing vital in our every-day life. The underlying data of such applications change very frequently, especially in the cloud environment. Many interesting events can be detected by discovering such data from different distributed sources and analyzing it for specific purposes (e.g., car accident detection or market analysis). However, several isolated events could be erroneous due to the fact that important data sets are either discarded or improperly analyzed as they contain missing data. Such events therefore need to be monitored globally and be detected jointly in order to understand their patterns and correlated relationships. In the context of current cloud computing infrastructure, no solutions exist for enabling the correlations between multi-source events in the presence of missing data. This paper addresses the problem of capturing the underlying latent structure of the data with missing entries based on association rules. This necessitate to factorize the data set with missing data.

The paper proposes a novel model to handle high amount of data in cloud environment. It is a model of aggregated data that are confidences of association rules. We first propose a method to discover the association rules locally on each node of a cloud in the presence of missing rules. Afterward, we provide a tensor based model to perform a global correlation between all the local models of each node of the network.

The proposed approach based on tensor decomposition, deals with a multi modal network where missing association rules are detected and their confidences are approximated. The approach is scalable in terms of factorizing multi-way arrays (i.e. tensor) in the presence of missing association rules. It is validated through experimental results which show its significance and viability in terms of detecting missing rules.

1. Introduction

Today every organization is facing the issue of handling a potentially large volume of data that come from multiple and distributed data sources. Data applications such as weather data, data streaming, sensor data, web logs and publish/subscribe applications, produce and consume large volume data. The underlying data of such applications is very dynamic and changes very frequently. Many interesting events can be detected by discovering such data from different distributed sources and analyzing it for specific purposes [1–3]. For instance, in the vehicle industry, the future driver assistance systems will need to discover, collect and analyze the dynamic information about cars environment and the drivers state. It will need to collect the information from various different sources which are distributed across different locations. For instance, the real time information from road sensors, radars, GPS, video and eye-tracking systems need to be instantly discovered, collected and analyzed. In addition, such systems require that the correlation of information from different sources is necessary in

[*] Corresponding author. Tel.: +44 1865 484572.
E-mail address: m.younas@brookes.ac.uk (M. Younas).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.11.002

order to get a consistent view of the real world and help drivers in making appropriate decisions and taking appropriate actions. However, the appropriate discovery, collection and analysis of such data are affected by various factors. For instance, reliable analysis of data collected can be affected by missing data. The loss of information and errors in the data collection process are the two main contributing factors to the missing data. The consequence of erroneous and missing data is that some important data sets may be discarded or improperly analyzed giving the incorrect information. Further, several isolated events may also have to be monitored globally and jointly detected in order to understand their patterns and correlation relationships, leading to adapt the system behavior and take appropriate actions considering a particular conjunction of events. Moreover, in a large network of computers or sensors, each of the components has some data about the global state of the system and much of the systems functionality relies on modeling the global state of the system which is constantly changing. It is necessary to keep the models up-to-date and seek for the incomplete information. Computing global data mining models, e.g. decision trees, k-means clustering in large distributed systems may be very costly due to the scale of the system and due to the communication cost, which may be high. The cost further increases in a dynamic scenario when the data changes rapidly [1].

The aforementioned issues will be studied in this paper in a new type of distributed environment, i.e. the cloud computing [4] by handling specific data that are the “association rules” [5]. Cloud computing is the most hyped trends for the last few years. However, to our knowledge, currently there exists no solution that enables the correlations between multi-source events in the cloud computing.

The paper proposes a novel model to handle a high amount of data in cloud environment. The proposed model takes into account aggregated data that are confidences of association rules. This paper addresses the issue of discovering and predicting the missing association rules from incomplete data on a cloud node, and by correlating them with data coming from other nodes. The proposed approach is distributed and is based on tensor decomposition [6]. The decompositions are applied to data arrays for extracting and explaining their properties. The proposed model deals with a multi modal network where missing association rules are detected and their confidences are approximated. For that, the association rules i.e. their confidences will be represented as arrays in each node, where the obtained arrays are incomplete and the results of correlation between the association rules with other nodes are represented by a tensor. In other words, our goal at first attempt is to capture the latent structure of the data via higher-order factorization in the presence of missing association rules. The second attempt is to recover the missing entries toward the distributed correlation of association rules over the cloud network.

The paper proposes a novel approach in order to discover the association rules locally on each node of a cloud and globally correlates the local results (over the cloud) to predict missing association rules.

Our salient contributions in this paper are:

- *Global correlation model*. To handle and analyze efficiently the high amount of data on the cloud network, we propose an aggregation data model related to confidences of association rules in presence of missing data.
- *A scalable algorithm*. We developed a scalable algorithms for tensor factorization to correlate the association rules in presence of missing rules over the cloud network and recover these missing entries.
- *Experiments*. To validate the obtained results, the distributed approach is evaluated with numerical experiments on simulated data sets in presence of incomplete and missing data.

The remainder of the paper is organized as follows: Section 2 presents a set of definitions and background that are used in the design of the proposed approach. Section 3 gives an overall picture of the proposed approach. Section 4 describes the local mining step with data representation and the applied algorithm. Section 5 presents the second step of distributed mining based on the tensor concept. Section 6 gives experimental results of the approach. Finally, we provide and outlook on future work and conclusion of the paper in Section 7.

2. Background

Notation

In data mining, the association rule is a popular and well research method for discovering interesting relations between variables in large databases [7]. Agrawal introduced associations for discovering regularities between products in large scale transaction data recorded in supermarkets. The deduced information can be helpful for decisions about marketing activities (see Table 2).

This section describes some basic definitions and concepts which are used in the design of the proposed model.

2.1. Itemsets and association rules

2.1.1. Definitions

Following Agrawal’ definition, the problem of association rule mining is defined based on itemsets.

Definition 1 (*Itemset*)**.** Let $I = \{i_1; i_2; \ldots; i_k\}$ be a set of k binary attributes called *items* and let $T = \{t_1; t_2; \ldots; t_n\}$ be a set of transactions, an item i_l is an attribute and $I \subset T$.

An itemset is characterized by the following concepts:

- *Support*. A support of an itemset I denotes $\text{supp}(I)$ is defined as the proportion of transactions in the data set which contains the itemset and is equal to the number of object containers.
- *Frequency*. The frequency of an itemset I is the probability that I occurs in the set of transactions T, which is denoted by $\text{Freq}(I)$ and is equal to $\frac{\text{supp}(I)}{\text{card}(T)}$ where $\text{card}(T)$ means the total number of transactions in T.

It is known that the itemset is frequent if its support is greater than or equal to a minimum threshold.

Theorem 1. *All itemsets form an ideal order in* $(2M, \subseteq)$ *(compared to the frequency constraint).*

We deduce that any subset of a frequent itemset is frequent, and any superset of an infrequent itemset is infrequent.

Definition 2 (*Association Rule*)**.** An association rule is expressed as: $R : X \longrightarrow Y$, with $X \in T$, $Y \in T$ and $X \cap Y = \emptyset$. The concepts related to a rule are:

- *Support*. The support of a rule is expressed by the amount of objects in T containing $X \cup Y$, ($\text{supp}(R) = P(X \cup Y)$). We measure the strength of an association rule by the confidence which is equal to the proportion of transactions containing X that also contain Y,
- *Confidence*. $\text{conf}(R) = \frac{P(X \cup Y)}{P(X)}$.
  Two types of rules emerge from the confidence measure: the exact rule if $\text{Conf}(R) = 1$ and rules of thumb if $\text{Conf}(R) < 1$.
- The “*lift*” rule measures the improvement provided by the association rule in relation to a set of random transactions (where X and Y are independent). It is defined by $\frac{P(X \cup Y)}{P(X)P(Y)}$. A “lift” greater than 1 indicates a positive correlation between X and Y, and thus the significance of the association.