Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 110 (2017) 428–433

Procedia Computer Science

www.elsevier.com/locate/procedia

International Workshop on Big Data and Networks Technologies
(BDNT-2017)

A Divided Latent Class analysis for Big Data

Abdallah Abarda[a,*], Youssef Bentaleb[a], Hassan Mharzi[a]

[a]*EECOMAS-LAb, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco*

Abstract

Statistical methods are a fundamental component in the big data environment. Among these methods: Latent class analysis (LCA), which is a subset of structural equation modeling, used to create classes in the case of multivariate categorical data. The use of this method to analyze massive data sets represents an expensive computational task. In this paper, we propose a Divide-and-Conquer approach for LCA model, the aim is to estimate the LCA parameters when this method is used for massive data sets. The performance of our approach will be verified by carrying out a numerical simulation.

Keywords: Latent class analysis, Massive data, Multivariate Categorical data, Statistical methods, Structural equation modeling

1. Introduction

The information technology revolution conducted researchers to find a solution for exploitation of massive data. Actually, the new technologies of information and communication generate Databases with Hundreds of fields, trillions of records, and terabytes of information. The massive data processing will likely play a crucial role in future society because it has an application in various fields such as marketing, health, education, customer services, sustainable development and transport.

In recent years, wide attention has been given to the problem of the use of statistical methods for massive data: Zhang and all[6] studied the Divide-and-Conquer kernel ridge regression (KRR), this approach partitions a data set of size N into m subsets of equal size, and computes an independent kernel ridge regression estimator for each subset, then averages the local solutions into a global predictor.

Li and all.[7] have proposed inferential tools for massive data to estimate a parameter. They employed a two-stage procedure.

[*] Corresponding author. Tel.: +212-61-040-9216
E-mail address: abdallah.abarda@uit.ac.ma

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.06.111

A similar solution has been proposed by Sunghae Jun and all. This involves dividing the initial population into M samples drawn by simple random sampling (until reaching the size of the population N), and then applying the linear regression method on each sample M[8].

In this work, we propose to use a Divide-and-Conquer approach for using Latent class analysis method for massive data, this approach partitions a data set of size N into B independent subsets of equal size. Our contribution focuses on two points: the first is the adaptation of Divided-and-Conquer approach to LCA method, the second is to provide a stop condition for the algorithm in order to minimize the number of subsets processed.

2. Formulation of Basic Latent Class Analysis

Let $\eta_{kc} = P(y_k = 1/c)$ be the conditional probability of the k^{th} variable $(k = 1, ..., V)$ being 1 given the c^{th} class $(c = 1, ..., C)$. Let y_i the response patterns or outcome vector $(y_i = (y_{1,i}, ..., y_{V,i})')$. The probability that an individual i $(i = 1, ..., N)$ has an outcome vector y_i is given by

$$\varphi(y_i) = \sum_{c=1}^{C} \rho_c \varphi_c(y_i) \tag{1}$$

where $\rho_c = P(c)$ is the class proportion and $\varphi_c(y_i) = P(y_i/c)$ is the conditional probability to have an outcome vector y_i given the c^{th} class, $(\sum_{c=1}^{c=C} \rho_c = 1)$.
Assuming the conditional independence of V variables conditional on latent class c, $\varphi_c(y_i)$ follows a Bernoulli distribution, and can be written as

$$\varphi_c(y_i) = \prod_{k=1}^{V} \eta_{kc}^{y_{k,i}} (1 - \eta_{kc})^{1-y_{k,i}} \tag{2}$$

The likelihood function for each individual is given by

$$\varphi(y_i) = \sum_{c=1}^{C} \rho_c \prod_{k=1}^{V} \eta_{kc}^{y_{k,i}} (1 - \eta_{kc})^{1-y_{k,i}} \tag{3}$$

By applying Bayes theorem, we can calculate the posterior probability

$$\psi_{ci} = \frac{\rho_c \varphi_c(y_i)}{\sum_{c=1}^{C} \rho_c \varphi_c(y_i)} \tag{4}$$

or

$$\psi_{ci} = \frac{\rho_c \prod_{k=1}^{V} \eta_{kc}^{y_{v,i}} (1 - \eta_{kc})^{1-y_{v,i}}}{\sum_{c=1}^{C} \rho_c \prod_{k=1}^{V} \eta_{kc}^{y_{v,i}} (1 - \eta_{kc})^{1-y_{v,i}}} \tag{5}$$

The EM algorithm[9] is often used to estimate LCA parameters, based on the maximum likelihood estimation (MLE). This algorithm is made of two important steps: the expectation step, denoted by E and the maximization step, denoted by M. The first step consists of calculating the expectation of the log-likelihood assuming that we have the information about classes. Two types of parameters are estimated: conditional probabilities (η_{kc}) and class proportions (ρ_c).
The number of parameters to be estimated for basic LCA is equal to :

$$p = C(V + 1) - 1 \tag{6}$$

3. Formulation of Divide-and-Conquer Approach for basic Latent class analysis (DACL)

Assuming that the samples were drawn randomly without replacement, and that each sample (sub-population) was fitted to the same number of class C as the entire population.