Multidimensional Mining over Big Healthcare Data: A Big Data Analytics Framework

Mario Bochicchio[1], Alfredo Cuzzocrea[2], Lucia Vaira[1],
Antonella Longo[1], Marco Zappatore[1]

[1]University of Salento, Lecce, Italy
{mario.bochicchio,lucia.vaira,
antonella.longo,marcosalvatore.zappatore}@unisalento.it
[2]University of Trieste and ICAR-CNR, Trieste, Italy
alfredo.cuzzocrea@dia.units.it

Abstract. Nowadays, a great deal of attention is being devoted to *big data analytics in complex healthcare environments*. *Fetal growth curves*, which are a classic case of *big healthcare data*, are used in prenatal medicine to early detect potential fetal growth problems, estimate the perinatal outcome and promptly treat possible complications. However, the currently adopted curves and the related diagnostic techniques have been criticized because of their poor precision. New techniques, based on the idea of *customized growth curves*, have been proposed in literature. In this perspective, the problem of *building customized or personalized fetal growth curves by means of big data techniques* is discussed in this paper. The proposed framework introduces the idea of *summarizing the massive amounts of (input) big data via multidimensional views* on top of which well-known *Data Mining methods* like *clustering* and *classification* are applied. This overall defines a *multidimensional mining approach*, targeted to *complex* healthcare environments. A preliminary analysis on the effectiveness of the framework is also proposed.

Keywords: Mining Big Data, Big Healthcare Data, Healthcare Systems.

1 Introduction

Big data analytics in complex healthcare environments (e.g., [13,14,15,16,17]) are of high interest at now, by following well-known principles of *big data management and mining* (e.g., [18,19,20]). Here, the main problem consists in devising models, techniques and algorithms focused to extract *useful knowledge* from enormous amounts of (big) data, with the goal of implementing so-called *big data intelligence*, i.e. deriving decisions, decision processes, guidelines and policies devoted to improve the target healthcare system (e.g., [13]). *Fetal growth curves*, which are a classic case of *big healthcare data*, are very important in prenatal medicine for fetal well-being evaluation. Indeed, they represent a mature and well-established practice to early detect potential fetal growth restriction, to estimate the perinatal outcome and promptly treat possible complications. The general idea underlying this test is very simple and effective: fetuses

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy. Copyright held by the author(s)

grow up showing a regular trend as a function of the gestational age. Therefore, their wellbeing can be assessed by tracking their sizes over the time and by comparing them with a reference growth curve known as "good". The implementation of the idea, based on ultrasounds pictures of the maternal abdomen, is quite simple, non-invasive and inexpensive.

In the clinical routine, fetal biometric parameters coming from this test are compared with a set of reference parameters, which are usually provided by the same test equipment. When results are too large or too small for the gestational age, they are classified as "potentially pathologic" and supplementary clinical tests are required/performed. A very problematic aspect in this practice is that several sets of fetal growth curves are reported in literature and the adoption of the right one is crucial to avoid errors (e.g., to avoid wrong classifications of fetuses as pathologic or non-pathologic) [5]. This is a hot topic for the obstetrics and gynecologist community [1], since the currently-adopted references lack of several mother-related aspects, such as ethnic group, food, drugs and smoke. Indeed, it has been recognized that these and other factors have a non-negligible influence on the actual growing trends of fetuses and, then, on the overall number of false-positives/negatives, and further unnecessary tests. In the current practice, failure rates as high as 46% are reported in literature [3], even considering the standard defined by the *World Health Organization* (WHO), so that, in several cases, it is hard to decide whether the fetus has to be considered pathologic or not. For this reason, *customized fetal growth charts* [1] have been proposed as an alternative to "literature-based" growth curves. The increasing acceptance of this best practice suggests for a new and ambitious perspective: the creation of an online service able to collect and analyze the world production of fetal growth data [2], in order to support obstetricians in the production of customized/personalized fetal growth curves. The clinical understanding of the phenomenon described by such a large amount of data is extremely intriguing, because of the underlying idea of finally grabbing a total understanding of the fetal growth processes. On the other hand, it is also challenging, due to both technical and medical reasons. These aspects are discussed in the remaining part of the paper. The main goal of the paper is that of assessing the feasibility of such online service. To this end, the paper proposes a *big data analytics framework for building customized or personalized fetal growth curves by means via innovative big data techniques*. The proposed framework introduces the idea of *summarizing the massive amounts of (input) big data via multidimensional views* [21] on top of which well-known *Data Mining methods* like *clustering* and *classification* are applied. This overall defines a *multidimensional mining approach*, targeted to *complex* healthcare environments. A preliminary analysis on the effectiveness of the framework is also proposed.

2 Building Customized Fetal Growth Curves by means of Big Data Analysis Techniques

In this Section, we provide the main contribution of our research, i.e. principles and definitions of a big data analytics framework for supporting multidimensional mining in complex healthcare environments. The idea of developing an online service to collect and analyze large datasets about maternal/fetal wellbeing and fetal growth, and supporting gynecologists and obstetricians in diagnoses of fetal growth restrictions has been considered valuable by several authors [1,3]. Actually, scaling this approach up to the worldwide production of fetal-maternal data could drive the medical community toward a deeper understanding of fetal pathologies, but several aspects have to be considered.

From a technical point of view, due to the volume of data to collect and analyze, the variety of descriptors associated to each mother/fetus and the velocity inherent to the phenomenon, Big Data techniques must be adopted. Indeed, every year there are about 160 millions of newborns in the world: on average this is equivalent to about 300 newborns per second. Considering that, according to international guidelines, for each healthy woman, about 10 clinical test (3 for growth tracking) are performed during pregnancy, and that for pathologic fetuses this number is significantly higher, the global production of data on the phenomenon can be estimated in a continuous stream of 1,000 – 10,000 new medical records per second, with an overall volume of 1 to 10 Petabytes per year. For the purposes of this paper, a more realistic scenario including 10% to 20% of fetal-growth data from at least two European countries is sufficient to test the proposed framework and tune it for further extension.

According to current methods adopted in the clinical practice for fetal-maternal wellbeing assessment, the main algorithms to analyze this stream would be based on: (*i*) *least-square method*, (*ii*) *multidimensional analysis*, (*iii*) clustering and classification techniques. The overall computational load is hard to estimate because of the problem is still under investigation, but *distributed approaches* and *parallelization techniques* are likely to be adopted. Moreover, the variety of data types (both structured and unstructured) is one of the main characteristics of this research field, because of the elements affecting fetal growth are not completely known and, every year, new variables come from the influence of new pathologies, medicines, therapies, pollutants etc. This heterogeneity is problematic to manage, but it is an important and unavoidable characteristic of the problem.

Referring to the algorithmic part, the possibility of constructing dynamic and customized fetal growth curves is mainly based on the following aspects:

- the application of *multidimensional analysis techniques*, which allow to both summarize the massive amounts of (input) big data via multidimensional and identify groups of patients (fetuses, in our case) who share similar growth patterns over the time;
- the possibility of searching for possible *correlation* of fetal growths vs parameters like ethnic group, maternal age, fetal gender, and so on.

The hypothesis is that fetuses at the same gestational age, with similar genetic make-up (e.g., ethnicity, familial aspects, and so forth) and in similar environmental conditions (e.g., food, smoke, drugs, and so forth), are subject to similar growth curves. This kind of fetuses will be referred, in the following of the paper as *Homogeneous Patient Groups* (HPG). This allows to identify patients who share common profiles in order to determine if a given fetus is potentially pathologic or not, when his/her growth parameters are different from those of the HPG to which he/she belongs. The membership to a specific group is established at run-time and a specific fetus can belong to one or more groups simultaneously, according to the analyzed features.

The diagnostic process is based on the following three main steps:

a) build the summarizing multidimensional view of the target experiment;
b) initially, the HPG of each mother/fetus is not known; it can be identified through anamnesis or specific tests and exams;
c) the wellbeing of each new fetus is assessed by comparing its actual sizes with the reference charts of the HPG identified by the previous step.

In terms of multidimensional analysis, patients can be represented as multidimensional points, and HPGs as regions, of a multidimensional space whose dimensions are all parameters affecting the fetal growth (ethnicity, maternal weight, height, familial aspects, foods, and so forth), and whose measures of analysis are the biometric parameters of the fetus. In this sense, step *b*) corresponds to identifying the patient's nearest HPG according to some distance measure, while step *c*) requires to compute the average size, the variance, and the corresponding percentile on a purposely defined (sufficiently-wide and updated) subset of elements of the same HPG. Moreover, HPGs can be periodically updated (e.g., every four to six months, considering that growth variations are not expected to emerge on shorter periods) by means of a suitable clustering algorithm. Considering that clustering algorithms are computationally intensive and cannot be repeated at the arrival of every new biometric fetal measure, a suitable classification algorithm must be exploited in order to decide to which precomputed clusters the new sample belong [10].

The innovative aspects of this method with respect to other approaches known in literature can be summarized as follows:

- HPGs are periodically updated rather than statically defined by means of standardized growth curves;
- reference growth curves are associated to each HPG (hundreds or thousands) rather than on few ethnic groups;
- fetal growth curves are continuously updated with the data coming from patients under examination, thus also revealing the long-term population trends in patient's growth.

3 Implementation, Preliminary Results and Discussion

We conducted some preliminary analysis to proof the effectiveness of our proposed framework, on the basis of real-life big healthcare data coming from a government/university research project. This Section describes the outcome of this task of our research.

In order to explore the implementation details of the proposed big data analytics framework for building customized fetal growth curves, we decided to collect and analyze an actual sample of fetal-maternal data coming from two facilities locate in the Apulia region in south Italy, namely: a university clinic, also involved in research about malformation and diabetes in pregnancy, and a general hospital. The facilities serve a basin of about 1.5 million citizens and assists more than 10,000 pregnant women per year. The sample, concerning about 500 pregnant women under assistance by 8 medical doctors, consists of a quite sparse table with about 2,500 records and 60 attributes grouped into 9 main categories (having obvious meaning):

- *Personal Data*;
- *Parity*;
- *Fetal Biometry*;
- *Diabetic Profile*;
- *Maternal Biometry*;
- *Familiarity*;
- *Glycemic Profile*;
- *Other Pathologies*;
- *Delivery Outcome*.

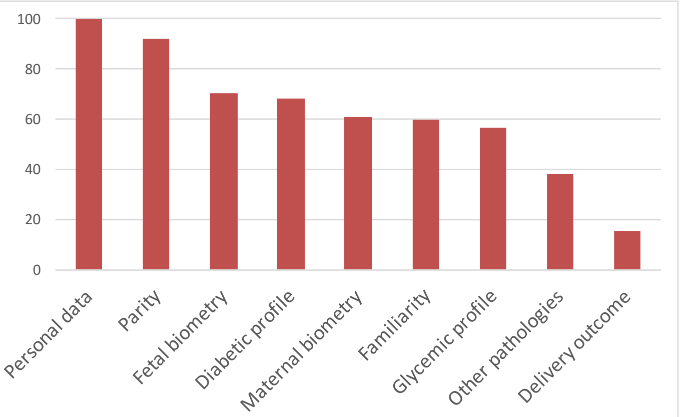

Fig. 1. Information Completeness.

For each category, the percentage of information completeness, defined as the number of not-null records over the total number of records, is represented in Fig. 1.

This collection permitted us to better understand the nature and the variability of data involved in maternal and fetal wellbeing monitoring. Moreover, it permitted us to define a set of *Dimensional Fact Models* (DFM) [23] able to describe a typical fetal-maternal test, along with its variable aspects. A simplified version of the main DFM is shown in Fig. 2.

According to the guidelines of the WHO on fetal growth curves, the available data sample has been also used to extract the reference curves for the target population to be analyzed. The process included a preliminary *normality distribution test* and a *linear regression*. This preliminary step was essential for a quantitative evaluation of the reduction of false positive/negative obtained with the new proposed method.

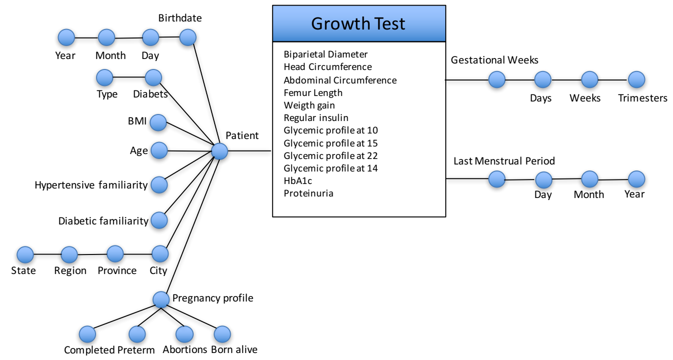

Fig. 2. DFM of a Fetal Maternal Test.

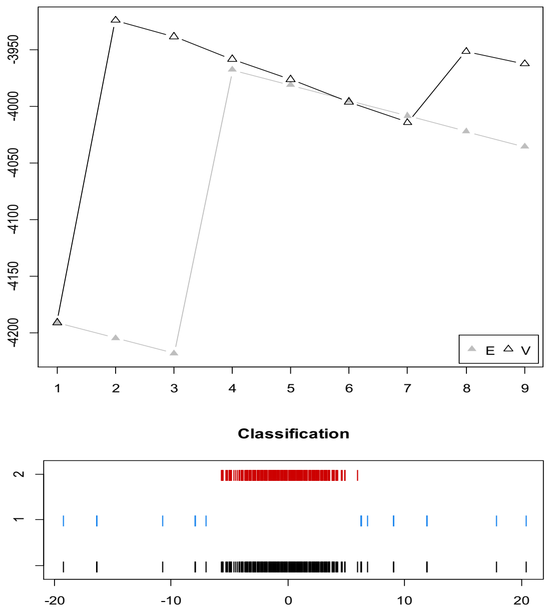

Fig. 3. Expectation Maximization Clustering Results.

For what concern the clustering analysis, the above-defined multidimensional model has been implemented on top of the OLAP server *Mondrian* [24] and two standard implementations of the density-based and EM clustering techniques have been provided by the *R environment* [25]. In our preliminary experimentation, we noticed that, while the EM algorithm converges on two overlapped clusters, no clear results come out from the density-based algorithm, probably due to the very non-homogeneous nature of the processed dataset. The results achieved by applying the EM algorithm are represented in Fig. 3 - up (Bayesian Information Criterion, used to estimate the number of clusters in the analyzed sample) and Fig. 3 - down (the original dataset and the two achieved clusters). Indeed, this approach can be improved by overcoming the well-known not-exciting performance of density-based clustering algorithms (e.g., [31]), for instance by adopting a kind of adaptive threshold like in [32].

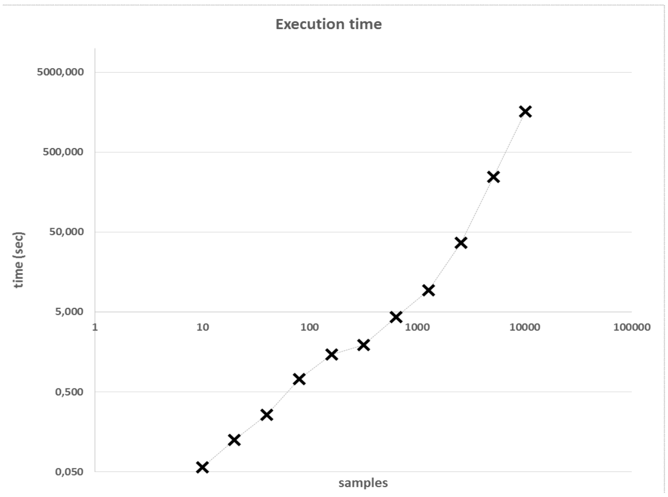

Fig. 4. EM Clustering: Execution Time vs. Sample Size.

Finally, in Fig. 4 it is reported the execution time of the adopted algorithm as a function of the problem size. This parameter is important to decide the maximum size of the clustered sample as well as how frequently it can be updated. The result shows that the execution time increases more than exponentially and that datasets of 5120 fetal sizes can be processed in about 4 minutes on a Pentium Core i5 @ 2.5 GHz, which is compatible with the discussed problem.

The application of these methodologies (i.e., multidimensional summarization and clustering analysis) confirms to us the effectiveness of our proposed framework in dealing with multidimensional mining in complex healthcare environments via big data analytics techniques.

4 Conclusions and Further Work

In this paper, we have introduced a big data analytics framework targeted to big healthcare data. The framework realizes a multidimensional mining approach for building customized or personalized fetal growth curves. The main idea consists in summarizing the massive amounts of (input) big data via multidimensional views on top of which well-known Data Mining methods like clustering and classification are applied. A preliminary analysis on the effectiveness of the framework has been also proposed.

Future work is mainly oriented towards extending our big data analytics framework by means of innovative computing metaphors such as *adaptiveness* (e.g., [29]) and *uncertainty* (e.g., [30]).

References

1. Gardosi J., Chang A., Kalyan B., Sahota D., Symonds E.M., "Customised antenatal growth charts", *Lancet 339(8788)*, pp. 283–287, 1992
2. Bochicchio M.A., Longo A., Vaira L., Malvasi A., Tinelli A., "Creating dynamic and customized fetal growth curves using cloud computing", in: *Proceedings of BIBE 2013*, pp. 1–4, 2013
3. Tinelli A. Bochicchio M.A., Vaira L., Malvasi A., "Ultrasonographic Fetal Growth Charts: An Informatic Approach by Quantitative Analysis of the Impact of Ethnicity on Diagnoses Based on a Preliminary Report on Salentinian Population", *BioMed Research International 2014 (1)*, Article ID 386124, 2014
4. Giorlandino M., Padula F., Cignini P., Mastrandrea M., Vigna R., Buscicchio G., Giorlandino C., "Reference interval for fetal biometry in Italian population", *Journal of Prenatal Medicine 3(4)*, pp. 62–68, 2009
5. Johnsen S.L., Wilsgaard T., Rasmussen S., Sollien R., Kiserud T., "Longitudinal reference charts for growth of the fetal head, abdomen and femur", *Eur J Obstet Gynecol Reprod Biol 127(2)*, pp. 172–185, 2006
6. Knorr-Held L., Best N. G., "A shared component model for detecting joint and selective clustering of two diseases", *Journal of the Royal Statistical Society 164(1)*, pp. 73–85, 2001
7. McLachlan G., Peel D., *"Finite Mixture Models"*, John Wiley & Sons, New York, USA, 2000
8. McLachlan G.J., Krishnan T., *"The EM Algorithm and Extensions"*, John Wiley & Sons, New York, USA, second edition, 2008
9. Bruno G., Cerquitelli T., Chiusano S., Xiao X., "A Clustering-Based Approach to Analyse Examinations for Diabetic Patients", in: *Proceedings of ICHI 2014*, pp 45–50, 2014
10. Pang-Ning T., Steinbach M., Kumar V., *"Introduction to Data Mining"*, Addison-Wesley, Boston, USA, 2006
11. Cerquitelli T., Chiusano S., Xiao X., "Exploiting clustering algorithms in a multiple-level fashion: A comparative study in the medical care scenario", *Expert Systems with Applications 55(1)*, pp. 297–312, 2016
12. Nithya N.S., Duraiswamy K., Gomathy P., "A Survey on Clustering Techniques in Medical Diagnosis", *International Journal of Computer Science Trends and Technology 1(2)*, pp. 17–22 , 2013
13. Sakr S., Elgammal A., "Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services", *Big Data Research 4(1)*, pp. 44–58, 2016
14. Lee C., Kim T., Hyun S.J., "A data acquisition architecture for healthcare services in mobile sensor networks", in: *Proceedings of BigComp 2016*, pp. 439–442, 2016
15. Barkhordari N., Niamanesh M., "ScaDiPaSi: An Effective Scalable and Distributable MapReduce-Based Method to Find Patient Similarity on Huge Healthcare Networks", *Big Data Research 2(1)*, pp. 9–27, 2015
16. Mezghani E., Exposito E., Drira K., Da Silveira M., Pruski C., "A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare", *Journal of Medical Systems 39(12)*, p. 185, 2015
17. Begoli E., Dunning T., Frasure C., "Real-Time Discovery Services over Large, Heterogeneous and Complex Healthcare Datasets Using Schema-Less, Column-Oriented Methods", in: *Proceedings of BigDataService 2016*, pp. 257–264, 2016

18. Cuzzocrea A., Saccà D., Ullman J.D., “Big data: a research agenda”, in: *Proceedings of IDEAS 2013*, pp. 198–203, 2013
19. Cuzzocrea A., “Analytics over Big Data: Exploring the Convergence of Data Warehousing, OLAP and Data-Intensive Cloud Infrastructures”, in: *Proceedings of COMPSAC 2013*, pp. 481–483, 2013
20. Yu B., Cuzzocrea A., Jeong D.H., Maydebura S., “On Managing Very Large Sensor-Network Data Using Bigtable”, in: *Proceedings of CCGRID 2012*, pp. 918–922, 2012
21. Gray J., Chaudhuri S., Bosworth A., Layman A., Reichart D., Venkatrao M., Pellow F., Pirahesh H., “Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals”, *Data Mining and Knowledge Discovery 1(1)*, pp. 29–53, 1997
22. Bailey T.L., “Fitting a mixture model by expectation maximization to discover motifs in biopolymers”, in: *Proceedings of ISMB 1994*, pp. 28–36, 1994
23. Golfarelli M., Maio D., Rizzi S., “The Dimensional Fact Model: A Conceptual Model for Data Warehouses”, *International Journal of Cooperative Information Systems 7(2-3)*, pp. 215–247, 1998
24. Mondrian – Pentaho Community, http://community.pentaho.com/projects/mondrian/, 2016
25. The R Project for Statistical Computing, https://www.r-project.org/, 2016
26. Agrawal D., Das S., El Abbadi A., “Big data and cloud computing: current state and future opportunities”, in: *Proceedings of EDBT 2011*, pp. 530–533, 2011
27. Xia F., Yang L.T, Wang L., Vinel A., “Internet of things”, *International Journal of Communication Systems 25(9)*, p. 1101, 2012
28. Dean J., Ghemawat S., “MapReduce: simplified data processing on large clusters”, *Communications of the ACM 51(1)*, pp. 107–113, 2008
29. Cannataro M., Cuzzocrea A., Pugliese A., “XAHM: an adaptive hypermedia model based on XML”, in: *Proceedings of SEKE 2002*, pp. 627–634, 2002
30. Cuzzocrea A., Kai-Sang Leung C., Kyle MacKinnon R., “Mining constrained frequent itemsets from distributed uncertain data”, *Future Generation Computer Systems 37(1)*, pp. 117–126, 2014
31. Aliguliyev, R.M., “Performance Evaluation of Density-based Clustering Methods”, *Information Sciences 179(20)*, pp. 3583–3602, 2009
32. Hassani M., Spaus P., Cuzzocrea A., Seidl T., “I-HASTREAM: Density-Based Hierarchical Clustering of Big Data Streams and Its Application to Big Graph Analytics Tools”, in: *Proceedings of CCGrid 2016*, pp. 656–665, 2016