MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220181

Applying Formal Concept Analysis for the Recognition of Infant Mortality Patterns

Lorenzo Marzano, Cristiane Nobre, Luis Zárate, Mark Song

Department of Computer Science, Pontifical Catholic University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Abstract

Infant mortality (IM), an index that corresponds to the number of deaths among children up to one year, is an important social indicator of a region. It generally reflects the conditions of socioeconomic development - in addition, the access and quality of resources available for maternal and child health care. Monitoring its magnitude, thus, can help in the definition of public policies for its confrontation. The main causes of IM can be also associated with biological, behavioral, and public health issues. In this work, implication and association rules based on Formal Concept Analysis are used to recognize patterns in births occurring in the state of Amapá (located in the Brazilian Amazon), where the index of infant mortality is more severe.

Keywords:

Public Health, Infant Mortality, Formal Concept Analysis.

Introduction

The infant mortality rate (*IMR*), the number of children who die in the first year of life considering the total number of children born alive in a given year and location, is a socioeconomic indicator widely used. Understanding its causes means raising possible solutions to improve the quality of life for the child and the mother.

Many factors can be associated with infant mortality such as biological (low birth weight, APGAR, maternal age, prematurity), socioeconomic (education, race, income), associated with health care (medical monitoring), behavioral (smoking, alcohol use), demographic (birth rate, fertility rate) [1, 2].

Brazil, being a country of continental dimensions, presents a great social inequality between regions. According to data from the Brazilian Institute of Geography and Statistics (IBGE), Amapá is the state with the highest infant mortality rate, with 23.45%. The state of Espírito Santo, for example, has an IMR considered low, more than twice as low as that of the state of Amapá for the same year, with 9.19%.

The objective of this work is to explore data made available by SUS (a health-related government agency) regarding births in the state of Amapá applying techniques of Formal Concept Analysis (*FCA*) to recognize patterns through the extraction of implication and association rules [3, 4].

FCA presents some fundamental concepts: formal context, formal concept, conceptual lattice, and association rules, which will be described. A formal context is mathematically defined as a triple $K:= (G, M, I)$ in which G is the set of objects, M the set of attributes and I is the incidence relationship, which can also be represented by a binary relationship between objects and attributes so that $I \subseteq G \times M$. From a formal context, algorithms can be applied to obtain a set of formal concepts, which aims to identify the set of attributes that delimit and characterize objects (*intention*) and the objects that share such attributes (*extension*). Mathematically, these entities are ordered pairs (A, B) where A and B are respectively subsets of the set of objects and the set of attributes. Given a set of objects $A \subseteq G$ from a formal context $K:=(G, M, I)$, it could be asked which attributes from M are common to all those objects. Similarly, it could be asked: "for a set $B \subseteq M$, which objects have the attributes from B?". These questions define the derivation operators, which are formally defined as:

$$A' := \{m \in M \mid gIm \ \forall g \in A\} \quad (1)$$

$$B' := \{g \in G \mid gIm \ \forall m \in B\} \quad (2)$$

Thus, the pair (A, B) is a formal concept if, and only if, $A = B'$ and $B = A'$. The set of formal concepts of a context $K:=(G, M, I)$ is denoted by $\beta:=(G, M, I)$. *Fom K or* β one can extract association rules which are dependencies between elements in a set. Let $K:= (G, M, I)$ be a formal context. An association between the attributes of M is a pair (X, Y), $X, Y \subseteq M$, receiving the notation $X \rightarrow Y$. The association rules, in turn, reveal frequent patterns of data.

Methods

This study was conducted to investigate the infant mortality rate (*IMR*) on Amapá (a brazilian state), located in the Brazilian Amazon, with the highest infant mortality rate. The procedures used in this study involves: data collection; analysis; exploration, selection, and binarization of attributes; creation of context, and rules extraction. The first step was to collect data from DATASUS , as well as to unify the database (DATASUS is the IT department of the Brazilian Unified Health System. An organ of the Secretariat for Strategic and Participative Management of the Ministry of Health, with the responsibility of collecting, processing and disseminating health information).

Subsequently, an analysis and exploration of the available attributes were carried out, with the objective of understanding and identifying those with missing or inconsistent data. The selection of attributes is also a critical step of this project since it is necessary to reduce dimensionality to work with the *FCA* approach. Once the main variables are selected, they must be binarized so that a context can be built. From the construction of the context, the *FCA* was effectively applied. The Unified Health System (SUS) has an extensive and varied database provided by the SUS Informatics Department (DATASUS). Due to the easy obtaining and availability of this information, the DATASUS databases were selected for this work. To acquire information related to infant mortality, it was necessary to work with two bases: one related to the Mortality System (SIM) and another related to the Live Birth System (SINASC). The two bases were unified, in order to obtain a list of children who survived and who did not survive the first year of life.

The infant mortality database in the state of Amapá, taken from DATASUS, has 13837 instances and 15 attributes, in addition to the classification, which defines whether the baby was born alive or not. The attributes are numeric and categorical, with missing data and other inconsistencies that must be analyzed. The attributes of this base are: mother's age, mother's education, number of live children, number of children dead, pregnancy, type of delivery, weight, sex, race, mother's marital status, APGAR1, APGAR5 and presence of a congenital anomaly.

The mother's age (Figure 1) is a numerical attribute, whose minimum and maximum values are 10 and 50 years, respectively. The data are diverse since there is at least one instance for each age within this range. In addition, the database does not present missing data for this attribute. The average distribution is approximately 25 years, with a standard deviation of 6.7 years.

The mother's schooling (Figure 2) corresponds to the years of studies completed and is an attribute that has 5 categories, in addition to some unclassified objects, represented by the question mark. The majority classification is between 8 to 11 years of schooling, representing 57% of the database. 107 instances were eliminated because they did not have the correct information.

The number of children (Figure 3), alive or dead, are discrete numerical attributes. In the case of the number

Figure 1 - Mother's age

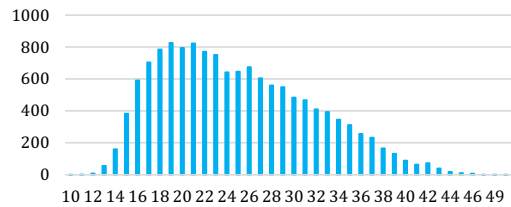
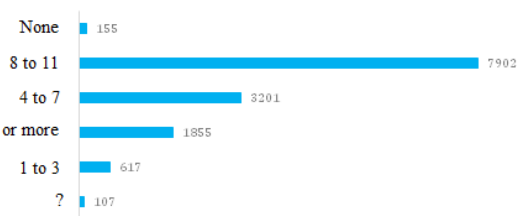


Figure 2 - Mother's schooling



of live children, the range varies from 0 to 15, containing samples throughout this range. There is also a progressive drop in the number of instances with the increase of live children so that a considerable portion of births corresponded to the first child. The number of dead children (Figure 4), in turn, presents values in the range of 0 to 15. However, the instance that has 10 dead children corresponds to an outlier. For this reason, this instance was eliminated from the study.

Pregnancy (Table 1) is a categorical attribute that indicates whether the mother was expecting a child, twins, triplets, or more. It is divided into three categories, 98% of which are single pregnancies. 15 instances were found with this missing data, so they were excluded from the database.

Figure 3- Living children

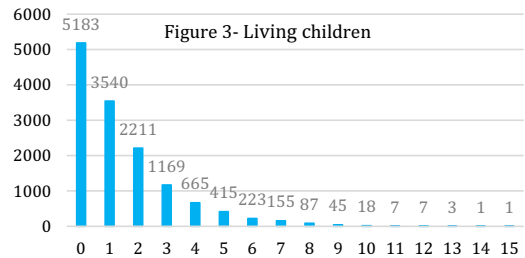


Figure 4 - Dead children

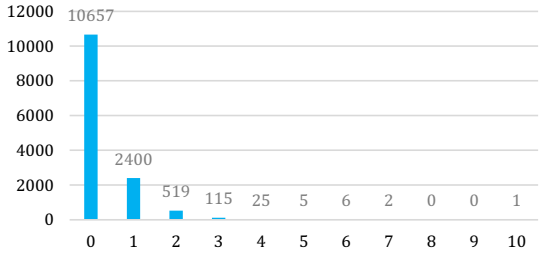


Table 1 - Pregnancy	
simple	13528
double	183
Triple or more	3
?	15

Gestation (Table 2) consists of the time, in weeks, that the pregnancy lasted. DATASUS already offers this discretized attribute, in 6 intervals of gestation time. More than 800 instances were discarded due to inconsistencies or lack of information.

Table 2 - Gestation	
Up to 22 weeks	17
22-27 weeks	77
28-31 weeks	137
32-36 weeks	1559
37-41 weeks	10434
More than 42 weeks	652
?	853

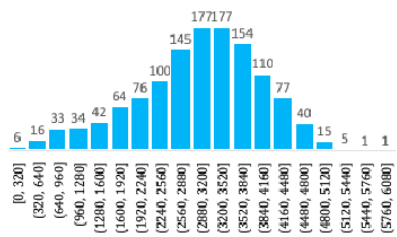
Child-birth (Table 3) can be classified as vaginal/normal or cesarean. A dozen copies were eliminated from the database because of missing data. It was also observed that approximately 2/3 of the cases corresponded to normal delivery, while 1/3 corresponded to cesarean.

Table 3 - Child-birth/Delivery	
normal	9042
cesarean	4675
?	12

Weight (Figure 5) is a numeric, continuous attribute that indicates the baby's weight, in grams, at birth. The average weight is 3.20 kilograms with a standard deviation of 568 grams. 27 instances were found with this inconsistent or missing information, being excluded from the base.

APGAR in the first minute (Figure 6) and in the fifth minute (Figure 7) are two measures widely used in the context of assessing the health of the newborn. It is an indicator developed by pediatricians, which consists of the assessment of five objective signs of the newborn: appearance, pulse, gesticulation, activity and breathing. The distributions of these two attributes are shown below. The instances whose APGAR1 and / or APGAR5 did not have information were eliminated.

Figure 5 - Weight



The sex (Table 4) and the baby's race (Table 5), the mother's marital status (Table 6) and the place of birth (Table 7) are all categorical attributes - their distributions are shown in the following tables.
The attributes "race" and "place of birth" presented some missing data, so that these instances were removed from the database.

Table 4 - Baby's sex	
female	6332
male	6499

Table 5 - Race	
yellow	6
white	1339
red	207
brown	10295
black	760
?	224

Table 6 - Marital Status	
Single	6295
married	1166
divorced	50
widowed	70
?	5303

Table 7 - Birth	
hospital	12317
health posts	197
home	172
other	34
?	111

Figure 6 - APGAR1

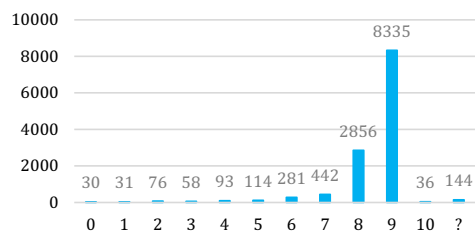
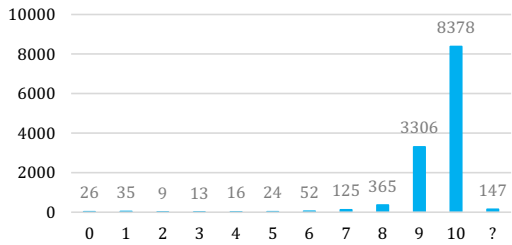


Figure 7 - APGAR5



The last attribute is the anomaly (Table 8), a categorical attribute that indicates whether the baby had any congenital anomalies. There is a major predominance of births without anomaly. Some instances were inconsistent and, therefore, were excluded from the database.

Table 8 - Anomaly

no anomaly	12165
some anomaly	73
?	111

The database, which initially contained 13837 instances, was reduced by 12% after exploring the attributes. With the remaining 12238 objects, we started to select the main attributes.

Two approaches to attribute selection are common in the literature: filter and wrapper [5]. The first approach selects the attributes in the pre-processing, disregarding the effects of the subset of selected attributes in the performance of the induction algorithm. The wrapper approach, in turn, does not present this disadvantage [6]. The algorithm is implemented in the data set, with a distinction between training and validation sets. Thus, that subset of variables with the greatest accuracy is chosen as the final set. This final set should be used in the sequence of the data mining tasks.

The selection of attributes was carried out through the WEKA *ExperimentEnvironment*, with the application of the *FilteredClassifier* classifier. The J48 algorithm (decision tree algorithm) was used to validate the subset of selected variables [7]. It was possible to observe five attributes in the rules of the decision tree: weight, pregnancy, anomaly, mother's education, and delivery. Thus, these five attributes were the selected variables. The model showed an accuracy of 99.1%, hitting 12,128 out of 12,238 instances. For the elaboration of the formal context, it was necessary to binarize the attributes. Therefore, the five attributes from the DATASUS database were restructured: weight, gestation, anomaly, mother's schooling, childbirth and classification. The binarization made the five selected attributes, in addition to the classification, become 21 binary variables for the generation of the context: gest_upto22 (gestation up to 22 weeks), gest_22to27, gest_28to31, gest_32to36, ges_37to 41, gest_42more, wei_upto1kg (weigth up to 1kg), wei_1to2kg, wei_2to3kg, wei_3kgmore, sch_none (mother's schooling = none), sch_1to3 (mother's schooling up to 3 years), sch_4to7, sch_8to11, sch_12more, cb_normal (childbirth normal), cb_cesarean (childbirth cesarean), ano_no (absence of congenital anomaly), ano_yes (congenital anomaly), c_alive (living child), c_dead (dead child).

The formal context is based on the binary variables. Table 10 shows part of the context obtained from Table 9 (based on some attributes of DATASUS database). As an example, 2 lines are presented. Each line represents information from the mother and child.

Table 9 – Database

gestation	weigth (grams)	schooling (years)	Childbirth Delivery	anomaly	classfication
< 22	930	8 to 11	normal	not	dead
32- 36	2120	8 to 11	cesarean	not	alive

Table 10 – Part of a Formal Context

gest_upto22	gest_22to27	gest_28to31	gest_32to36	gest_37to 41	gest_42more	wei_upto1kg	wei_1to2kg	wei_2to3kg	wei_3kgmore	sch_none	sch_1to3	sch_4to7	sch_8to11	sch_12more	cb_normal	cb_cesarean	ano_no	ano_yes	c_alive	c_dead
x						x							x		x		x			x
			x					x					x			x	x		x	

Results

When applying the *FCA* techniques, it was possible to find 1345 concepts. Based on these concepts, 259 implications were obtained in the following format *Premise => Conclusion*. 40 rules did not contain objects to support them. 19 implications were found with more than 100 objects (records) supporting them with 100% confidence (Table 11).

Nine hundred and fifty three (953) association rules with *confidence ≥ 80%* were found. Association rules are true only for some portion of the totality of objects, which are covered by the premise - different from implications, non-strict rules are allowed, that is, rules, for which, if the premise is valid, the conclusion is not necessarily valid. Table 12 shows some association rules found in the format: <Number of objects, for which premise holds> Premise [Rule confidence] => <Number of objects, for which premise and conclusion hold> Conclusion. It is interesting to note that many of the association rules generated were extremely reliable, being rounded up to 100%.

Table 11 – Implications

1 wei_2to3kg sch_8to11 cb_normal ano_no => **c_alive**;
2 sch_12more cb_cesarean => ano_no
3 gest_37to41 sch_12more cb_normal => **c_alive**
4 gest_37to41 sch_4to7 cb_cesarean ano_no => **c_alive**
***
14 wei_1to2kg cb_normal c_alive => ano_no
15 gest_32to36 wei_1to2kg c_alive => ano_no
16 wei_2to3kg sch_1to3 => **c_alive**
17 sch_none ano_no => **c_alive**
18 gest_42more wei_2to3kg => ano_no **c_alive**
19 gest_37to41 wei_2to3kg sch_1to3 ano_no => **c_alive**

Table 12 – Association rules

<12238> { }[99%] => <12165> ano_no
<12123> **cb_alive** [99%] => <12061> ano_no
<12165> ano_no [99%] => <12061> **cb_alive**
<9926> gest_37to41 [100%] => <9902> **cb_alive**
<9874> gest_37to41 ano_no [100%] => <9855> **cb_alive**
<9902> gest_37to41 **cb_alive** [100%] => <9855> ano_no
<12061> ano_no **cb_alive** [82%] => <9855> gest_37to41
<8437> wei_3kgmore [100%] => <8417> **cb_alive**
<8399> wei_3kgmore ano_no [100%] => <8382> **cb_alive**
<8417> wei_3kgmore **cb_alive** [100%] => <8382> ano_no
<8065> cb_normal [99%]=> <8020> ano_no
<7987> cb_normal **cb_alive** [100%] => <7948> ano_no
<8020> cb_normal ano_no [99%] => <7948> **cb_alive**
<7323> gest_37to41 wei_3k,ore [100%] => <7309> **cb_alive**
<7292> gest_37to41 wei_3kgmore ano_no [100%] => <7280> **cb_alive**
<7309> gest_37to41 wei_3kgmore **cb_alive** [100%] => <7280> ano_no
<8382> wei_3kgmore ano_no **cb_alive** [87%] => <7280> gest_37to41
<7237> sch_8a11 [99%] => <7190> ano_no
<7168> sch_8a11 **cb_alive** [99%] => <7129> ano_no
<7190> sch_8a11 ano_no [99%] => <7129> **cb_alive**

Discussion

It is possible to observe many implications that present the classification "alive" as a consequence of the premise (Table 11). Of these rules, the pregnancy had a period always greater than 32 weeks, which indicates that longer periods of gestation are a relevant factor for the child's survival. It is also possible to highlight that the child's weight at birth appeared in many of these rules. With only two exceptions (rules 14 and 15), the weight tends to be more than 2 kilograms for the child to survive in the first year of life. It was also observed that there was no congenital anomaly in the premises of these main rules that "alive" was a consequence.

Rule 1 (Table 12) shows the significant number of children who were born without congenital anomalies. Therefore, for this database, 99% of the children did not have a congenital anomaly. Still on this variable, rules 2 and 3 express a close relationship between the absence of anomaly and the survival of children, with 99% of children without anomaly surviving and 99% of surviving children having no anomaly. It is also observed that pregnancy of 37 to 41 weeks leads, with practically 100% confidence, to a healthy child who will survive his first year of life. This same observation can be made when it comes to the weight variable, especially when the child is born with more than 3 kilograms.

Conclusions

Concerning infant mortality, the implications and association rules showed mainly the influence of biological characteristics, such as weight and gestation time, as determinants for the health and survival of children up to one year of age. As a proposal for future work, a study focused on the instances whose classification is "dead" is suggested, aiming to identify possible variables that are directly linked to this problem. In this work, many rules related to surviving children were observed. Although it is important to identify factors that are associated with the health of babies, it is also essential to identify and understand the factors related to their death. Another proposal is to explore more regions of Brazil, or even to focus on micro-regions within the state, to identify different trends in a single region.

Acknowledgements

The authors thank FAPEMIG (CEX-APQ-00997-15), CNPq (404431/2016-0) and CAPES (1196329) for the financial support.

References

[1] SRIDEVI, S.; NIRMALA, S. ANFIS. *Based decision support system for prenatal detection of Truncus Arteriosus congenital heart defect*. Applied Soft Computing, V. 46, 2016. Pages 577-587. ISSN 1568-4946.

[2] TESFAYE, Brook; ATIQUE, Suleman; ELIAS, Noah; DIBABA,Legesse; SHABBIR, Syed-Abdul; KEBEDE, Mihiretu. *Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach.* Computer MethodsandPrograms in Biomedicine, V. 140, 2017. Pages 45-51. ISSN 0169-2607.

[3] B. Ganter,and R.Wille. *Formal Concept Analysis: Mathematical Foundations.* Springer-Verlag New York, Inc., 1997. ISBN 3540627715.

[4] C. Carpineto,and G. Romano. *Concept Data Analysis: Theory and Applications*. John Wiley & Sons, 2004.ISBN 0470850558.

[5] KOHAVI, R.; JOHN, G. H. *Wrappers for feature subset selection*. Artificial intelligence, Elsevier, v. 97, n. 1, p. 273–324, 1997.

[6] JOHN, G. H. et al. *Irrelevant features and the subset selection problem*. In: ICML. v. 94, p. 121–129. 1994.

[7] OLIVEIRA, José G.; NORONHA, Robinson Vida; KAESTNER, Celso A. Alves. *Método de Seleção de Atributos Aplicados na Previsão da Evasão de Cursos de Graduação.* Revista de Informática Aplicada, v. 13, n. 2, p.54-67. 2017.

Address for correspondence

Mark Alan Junho Song - song@pucminas.br