Exploring Social Data to Understand Child Labor

Diego C. Rodrigues, David N. Prata, and Michel A. Silva

***Abstract*—Child labor is an issue of utmost importance to the world. According to the ILO (International Labour Organisation) nearby 218 million children between 5 and 17 work in the world, of which 50% holds hazardous work, The question rises up on how to locate and understand the factors about those families that generates the indices of child labor, and which properties are important to analyze. By the use of data mining techniques to discover valid patterns among Brazilian social data bases, we evaluate child labor in the federated state of Tocantins. This work has the purpose towards uncovering the deterministic factors for the practice of child labor and their relationships among financial, educational, cultural and social indicators; generating information that was not aware provided by data bases feeders, being hidden among records.**

***Index Terms*—Data mining, social data, child labor, welfare**.

I. Introduction

The Social Program System called Unique Registry (UR) from Brazilian Federal Government is an instrument to identify and qualify low-income families defined as those who have: monthly income of up to half of the Country basic salary per person or total monthly income of up to three basic salaries.

The registry database (UR) allows understanding the reality of these socio economic families, bringing information around the core family as for example, the characteristics of the residences, sorts of access to essential public services and also information about each of the components of the family.

The federal government has the intention to consolidate through computerized systems the data collected in the UR to spread out knowledge about low-income families to the public power. Based on this information, the public power can be able to formulate and implement specific policies that contribute to reducing social vulnerabilities that these families are exposed to. Currently, the UR has over 21 million registered families.

The UR is coordinated by the Ministry of Social Development and Fight against Hunger (MDS) and must necessarily be used for selecting beneficiaries of social programs from the federal government, such as the Family Fellowship (FF) program. UR can also be used by federated states and local governments to get the diagnosis of socio economic families' registry, enabling the development of local social policies.

Applying data mining techniques in order to discover valid patterns and knowledge is not a trivial task because of the large amount of data and attributes available in the UR. In order to analysis the UR data, we apply mining techniques in context to discover patterns related to child labor in the federated state of Tocantins. The data base was divided into two data sets, combining families with 168 attributes and people with 214 attributes, totaling a data set of 892,422 records to qualify this diverse information.

According to the IBGE (Brazilian Institute of Geography and Statistic) Census 2010 (another Brazilian social registry database) the northern region of Brazil is the number one in rates of child labor. From this point, some issues need to be identified, as the necessity to locate these evidences from IBGE families' conditions to check these cases so that the government actions have ability to reach and support these people. However, a problematic fact comes up, how to know if indeed this family has all the features necessary to have the government attention? Despite the amount of data collected quarterly by the federal government there may be flaws registration or registrations with omissions.

According to Census 2010 (IBGE), Tocantins federated state has a range estimation of 40 thousand households with indices of child labor. But, in fact, the Unique Registry demonstrates an amount of 5,250 families with child labor. This inconsistency induces the government to conduct an active search to locate and clarify the families whose are unsuitable enrolled in social programs and those who are not enrolled, but should be, and make the required corrections. The question rises up on how to locate and understand the factors about those families that generates the indices of child labor, and which attributes are really important?

This work has the purpose towards answering the following questions: "What are the deterministic factors for the practice of child labor?" and "What are the real relationships among financial, educational, cultural and social indicators?"

In fact, there is a strong tendency to associate infancy work with family financial conditions of the child, but is it really a deterministic indicator for this condition? And, how about considering others indicators?

II. Data Organization

The Unique Registry is maintained for over 15 years by the Brazilian government. This database has the function to register all Brazilian families who are at social risk. It stores a data set of families and their members, creating a powerful set of data with many types of potential information. The database is divided into two groups, the families and the

Manuscript received September 8, 2013; revised November 10, 2013.

Diego C. Rodrigues is with the Department of Labor and Social Welfare of the State of Tocantins (e-mail: diego.msn@globo.com).

David N. Prata is with Federal University of Tocantins, CO 77.001-009 BRA. He is coordinating the Master Degree Program of Computational Modeling, Palmas-TO (e-mail: ddnprata@uft.edu.br).

Michel A. Silva is with the Computational Modeling in Federal University of Tocantins, CO 77.001-009 BRA (e-mail: tensoja@gmail.com).

individuals. The data of the families are all information regarding the type of housing, income and social information, and about the family economic condition. The data concerning to people are their features and personal data, such as school, social and financial information. Thus, the government generates a complete registry of families and their members. For our study, we compose each family data with their respective people data. After the use of selection attribute algorithms and specialists interviews, we elect the following indicators to data mining: family income, education, sex, family that receives assistance from government for child study.

III. Knowledge Discovery

Data Mining is part of a larger process of research called Knowledge Discovery Database (KDD) defined as the exploration and analysis, automatic or semi-automatic, large quantities of data in order to discover meaningful patterns and rules [1].

The Data Mining techniques are becoming increasingly popular as tool for KDD to find hidden information necessary for decision making. However, this approach is difficult to apply because of its interdisciplinary abilities to combine different methods and techniques such as database, statistical methods, neural networks, genetic algorithms, machine learning, natural language processing and other areas of study.

The work data are not always in perfect condition to start the mining process. Data often have multiple sources or, are incomplete or, have noises. These unclear data setup need a one-step data formulation called pre-processing, including activities such as cleaning, integration, selection and transformation of data [2].

Once the preprocessing data is finalized, it could be loaded into data mining software, as Weka (Waikato Environment for Knowledge Analysis) [3]. Weka includes a series of algorithms for data formatting, machine learning algorithms and results validation, written in Java programming language [4] (it's open source code is available on Internet.)

Among many techniques available for data mining, the classification task could be considered the most suitable technique for study purposes. The data analysis objective was to classify social indicators with the intuition to discover the hidden patterns among them related to child labor.

IV. Decision Trees

The task of classification used a special type of artificial intelligence known as decision trees. Based on the records of the training set (the data used to train the algorithm), a tree diagram is designed and mounted. The design of the decision tree allowed us to classify the rough sample in branches disallowing the necessity to test all the values of its attributes which makes this type of sorting one of the best performance searches and most used algorithm in data mining [5].

In a decision tree, knowledge is represented by nodes and arcs in a hierarchal frame. Each parent node may lead the search to one of his sons. Hence, from the root down toward the leaves of the tree, the system configuration is modeled, and therefore the associated behavior.

The quality of the trees generated by the system was analyzed by a group of social assistance experts, which evaluated the classification of indicators as to their relevance regarding the potential cause of child labor.

V. Methodology

The first stage of the work consisted of examining data mining algorithms and chose one that could find patterns among attributes. The J48 algorithm was appointed by the IEEE International Conference on Data Mining (ICDM) [6] as the most promising algorithm for generating decision trees and one of the most popular approaches in data mining, Table I.

Decision tree is a classic technique to represent information from machine learning algorithms, and offer a fast and powerful method to express data structures.

TABLE I: Pseudo Code of J48 Algorithm

1. Check if algorithm satisfies termination criteria
2. Computer information-theoretic criteria for all attributes
3. Choose best attribute according to the information-theoretic
4. Create a decision node based on the best attribute in step 3
5. Induce (i.e. split) the dataset based on newly created decision node in step 4
6. For all sub-dataset in step 5, call J48 algorithm to get a sub-tree (recursive call)
7. Attach the tree obtained in step 6 to the decision node in step 4
8. Return tree

In this work we used an implementation of the generating algorithm of decision trees using the Waikato Environment for Knowledge Analysis (Weka) tool [3]. This version of the algorithm known as J48 is a WEKA implementation of the C4.5 algorithm [5]. The algorithm uses a greedy technique to induce decision trees for classification and uses the reduced error pruning technique. The Weka ease of use and the amount of resources offered by the tool led to its adoption instead of a new implementation of the algorithm for tree generation.

A. Preprocessing

Data only have quality if they meet the requirements of the intended use. There are many factors that guarantee the quality of the data, including the accuracy, completeness, consistency, timeliness, credibility and interpretability [7].

To ensure the quality measures of data preprocessing, some steps need to be completed, as following.

B. Data Integration

The database from UR was stored in different tables. To help reducing redundancies and inconsistencies in the working data set we performed an integration of the data which used the sample code as a key link among Excel spreadsheets. The final data set was created in a new file type CSV (comma-separated values). Redundant data were

grouped depending on the value or disposed in relation to the unique identifier of the table, avoiding inconsistencies in the data set.

C. Data Cleaning

At this stage some routines were performed to try to ensure the quality of data, such as, the missing data were replaced by a global constant indicated by "?". Insofar, the algorithm could handle the data gaps to not mislead the results.

D. Data Reduction

After cleaning the data, the final data have a set of attributes from the original set been reduced by performing a reduced dimension wherever irrelevant attributes or weakly redundant data could be detected and removed.

For this task we employed CfsSubsetEval algorithm to assess the value of a subset of attributes, considering the predictive ability of each individual feature, along with the degree of redundancy among them. Preferred subsets have highly correlated features among their classes, but having low intercorrelation [7].

For this work, the combination of BestFirst (search method) and CfsSubsetEval (attribute evaluator) is as efficient as the best techniques for variable selection, genetic algorithm and simulated annealing algorithm, being much faster [8].

To evaluate the attributes, we compared values using the heuristic merit of each relationship formalized by the equation in Formula I.

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Formula I. Equation to formalize the heuristic merit.

The final formula of merit uses the Pearson correlation between a variable composite (sum or average) and a target variable (the class in question) [6].

The algorithm of the Weka CfsSubsetEval ran with the initial set of data as input. From the 92 starting attributes, the base was reduced to 35. After evaluation of experts in social assistance, this number was further reduced to a total of 5 attributes considered essential for modeling the problem. The list of attributes to the respective scores of merit is shown in Table II.

TABLE I: SELECTION OF ATTRIBUTES USING CFSSUBSETEVAL ALGORITHM OF WEKA TOOL

Attribute Rated	Merit	Selected Attributes
Child Sex	0.01	• Index of Child Labor
Person with Benefits from Family Fellowship (FF) Government Program	0.084	• School Attendance Frequency • Family Income
School Attendance Frequency	0.018	• Person with Benefits from Family Fellowship (FF) Government Program • Index of child labor
Family Income	0.084	• Person with Benefits from Family Fellowship (FF) Government Program
Index of Child Labor	0.018	• Child Sex • School Attendance Frequency

E. Transformation of Data

The decimal numerical values were replaced by causing them to be correctly interpreted by Weka tool. Dates are formatted in "dd/mm/yyyy" standardizing the field type.

VI. EXPERIMENTATION

The first step was the identification of nonconformity for data quality through missing data in UR database. This evidence required a data refinement in all of the 840,000 records by verifying the existence of empty fields and null values for the 5 selected attributes (child gender, income family, school attendance frequency, child labor index, existence of family grant assistance from government programs), in order to adjust the database to be interpreted by the software package Weka.

The second step was the selection and joint of data by the use of a Structured Query Language (SQL) tool to support data manipulation through the selection of the records without missing values on their fields. This step allowed validating and preparing all data to be exported to CSV (Comma - separated values) file format data set.

The third step was to identify and group the most relevant indicators of the child labor. The objective was to create a dataset from the UR database represented only by the records in which the indicative of child labor are or positively or negatively marked, leaving out the null and missing values. The result was a total of 299,614 records with these markings. This step allowed to assess the reliable data context and to apply the knowledge discovery techniques.

The Fourth step was to validate the data refinement of the UR database with the government social experts. All the selected indicators were placed to the Ministry of Social Development and Fight against Hunger team, in order to evaluate the selected dataset and validate them. So the data set was secured as confident for the ongoing steps to import to Weka tool.

After the conclusion of the preprocessing step, the dataset was imported to the Weka tool in order to be classified by the use of J48 algorithm, generating a decision tree, Fig 1.

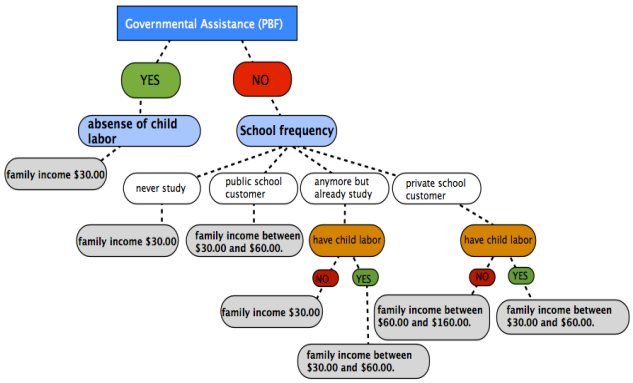

Fig. 1. Decision tree generated by the J48 algorithm from Weka.

VII. RESULTS AND DISCUSSION

The result of the tree generated by the algorithm was

presented to the government social assistance team from Tocantins federated state which evaluated and interpreted the outcomes in conjunction to the data mining experts.

The attribute used as basis for checking the data related to child labor was Person with Benefits to Family (PBF) of Social Assistance from Government Programs, which found the incidence of child labor and the relationship with other attributes (household income, incidence of child labor, child sex, school attendance frequency). Social factors like educational incomes for the incidence of child labor were analyzed.

For "no incidence of child labor" (the left part of the tree, Fig. 1), we can see the first node YES, which means that the child's family receives government assistance from social program FF (Family Fellowship). This means that for the data set comprising the base of the UR Tocantins, 2013 version, the child labor was eradicated from families whose receive benefits from PBF. Looking from a social point of view, this information is very important considering the social risk of child labor for all these families who falls into a situation of low-income families. This is and indicative that the government assistance programs could favor the eradication of child labor.

The income factor was another fairly observed field in the tree generated. The ratio of family income attribute where there was no case of child labor is in a range of up to $30.00. On the other hand, we can observe the nodes of the tree with incomes factors rates between $30.00 and $70.00, there is child labor cases, leading us to argument that the attribute income is not conclusive to child labor. This information need more profound study and extend research to understand the underline motivation about these families behavior about child labor. Otherwise, this evidence should supply the government under which actions to take to campaign against this practice of labor.

The attribute school frequency attendance has 4 possible values for the child's schooling (never studied, study in a public school, anymore but already studied, private school). We can note that these school question factors had a greater relationship with the occurrence of child labor. For example, for the attribute never studied there is no occurrence of child labor. But, for the child who already study and is not studying anymore has occurrence of child labor. We can observe that the work of child is considered as child labor, until they have 17 years old. In this case, we can conclude that if these children already study, therefore, these children are the older ones. For the child marked as public school and the ratio of family incomes between $30.00 and $60.00 there is no child labor occurrences.

An interesting point for the attribute school frequency attendance is the occurrence of child labor when children study in private school. For this node we can be observe that there is incidence of child labor for the incomes between $30.00 and $60.00 and the index was marked as negative when the incomes are between $60.00 and $160.00. This result hints us to consider some hypotheses, because of the child is in a private school. We can suggestion some meanings, such as the children are working to keep their study of best quality, or the child receives a scholarship to study. Another explanation comes so far from the belief of some families that the elementary public schools in Brazil have lower quality, despite that this is nowadays a questionable issue. Even though some families do not need government aid compared to families who are in a worse finance situation, these children families could be looking for a best quality of study.

In the least, the results and the discussion suggest that governments should employ researches of data mining to take better government social assistance decisions.

VIII. Conclusion

Child labor is an issue of utmost importance to the world, according to the ILO (International Labour Organisation) nearby 218 million children between 5 and 17 work in the world, of which 50% hold hazardous work.

Asia, Africa and South America has the highest rates of child labor, however, the rich and developed countries are not outside of that context.

According to UNICEF (United Nations Foundation for Children), India has approximately 14% of children between 5 and 14 years old are involved in child labor activities. Most of them work in family homes and subcontractors. Government supervision passes away from this kind of work, and these children are subjected to exploitation. Today, the Asiatic continent corresponds to the highest rate of children working in the world.

This issue is of concern, because these children are faced with any kind of work, seeking for their sustenance and family, unaware of protection that a home could offer.

In Brazil, policies to combat child labor are claimed by the federal government. Applying knowledge discovery techniques and standards as a strong alliance to social programs to identify, locate and understand children who are in this social risk is paramount in a nation with the 5th largest area in the world. Data mining could enable an active search to strongish the social activities by the supply of knowledgeable information and support decision actions to identify and combat the resulting facts for this world unsolved social problem.

For future works, the outcomes evoke the accomplishment of others researches which could drive the comparison data between timelines, e.g. before and after government social assistance. Other researches could consider comparing others Brazilian federated states towards child labor behaviors.

Acknowledgment

We would like to thank the social professional workers, Euvanilde Silva Brito, Carmen Lucia Vendramini, Daniela Nunes Alves Queiroz, Aurora Moraes, Gildeth Evagelista de Macedo and the Secretary of Labor and Social Welfare of the State of Tocantins, for their dedication and effort to make this work. We would like to also thank Professor Patrick Letouze for his contribution.

References

[1] G. S. Linoff and M. J. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 3rd ed., Wiley, 2011.
[2] S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi, "On the computation of

multidimensional aggregates," in *Proc. 1996 Int. Conf. Very Large Data Bases (VLDB'96*), Bombay, India, pp. 506-521, Sept. 1996.
[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, iss. 1, 2009.
[4] Oracle. (August 2013). The Java Programming Language and the Java Platform. Available: http://www.oracle.com/technetwork/topics/newtojava/downloads/index.html
[5] X. D. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, N. Angus, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl, Inf. Syst*, vol. 14, no. 1, pp. 1-37, 2007.
[6] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D diss., Waikato University, Hamilton, NZ, 1998.
[7] I. V. Tetko, V. P. Solov'ev, A. V. Antonov, X. J. Yao, J. P. Doucet, and, B. T. Fan, F. Hoonakker, D. Fourches, P. Jost, N. Lachiche, and A. Varnek, "Benchmarking of linear and nonlinear approaches for quantitative structure−property relationship studies of metal complexation with ionophores," *Journal of Chemical Information and Modeling* 2006, vol. 46, no. 2, pp. 808-819.
[8] O. E. S. Paulo. (August 2013).Worldwide, 218 million children work. http://www.estadao.com.br/noticias/internacional,218-milhoes-de-crianc as-trabalham-no-mundo-calcula-oit,10222,0.htm
[9] IBGE. (August 2013). 2010 census child labor. http://censo2010.ibge.gov.br/trabalhoinfantil
[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

Diego C. Rodrigues was born in Petrolina, Pernambuco on September 25, 1986, trained in analysis and development of systems for the State of Tocantins (2010), and graduate degree in project management from the University of Jacarepagua Rio de Janeiro (2013), course development ADOBE rich applications, Special Student MSc computational Modeling Systems acting professionally as a project manager and systems analyst for the department of work and welfare of the state of Tocantins in government social project, deploying and developing computer systems, acting as teacher Federal University of Tocantins (UFT) as major field of study data mining on social bases.

David Prata was born in Goi ânia, Brazil on September 18, 1965. Dr. Prata completed his Bachelor of Computer Science in 1992. Then on, he went to complete his specializing in Academician. He worked as a system analyst to Tocantins Government, being in charge for the accountability and financial systems. Later, he successfully completed his Master Degree in Computer Science from Campina Grande Federal niversity, with application research in education in 2000 year. He coordinated graduate and undergraduate courses in computer science at Alagoas Faculty in Maceio, Brazil. He was allotted to Federal University of Alagoas in 2006. Then, he moved to Federal University of Tocantins. His doctoral was developed in part at Carnegie Mellon University, USA, completed in 2008. He is currently coordinating a Master Degree in Computational Model. His research interests are education and ecosystems.

Michel A. Silva graduated in Computer Science from the Federal University of Tocantins (2008). Currently pursuing a masters of computer modeling system by the Federal University of Tocantins and works as a legislative assistant esp. programming - Legislature of the State of Tocantins, Brazil, mainly in the following areas: development, php, java, rails, c + +, usability, information architecture, data mining.