# Applicability of Data Mining Technique using Bayesians Network in Diagnosis of Genetic Diseases

Hugo Pereira Leite Filho
University State of Goiás: Information Systems:
Ceres, Brazil

***Abstract*— This study aims to identify a methodology to aid in the identification of diagnosis for chromosomal abnormalities and genetic diseases, presenting as a tutorial model the Turner Syndrome. So, it has been used classification techniques based in decision trees, probabilistic networks (Naïve Bayes, TAN e BAN) and neural MLP network (Multi-Layer Perception) and training algorithm by error retro-propagation. Described tools capable of propagating evidence and developing techniques of generating efficient inference techniques to combine expert knowledge with data defined in a database. We have come to a conclusion about the best solution to work out the show problem in this study that was the Naïve Bayes model, because this presented the greatest accuracy. The decision - ID3, TAN e BAN tree models presented solutions to the indicated problem, but those were not as much satisfactory as the Naïve Bayes. However, the neural network did not promote a satisfactory solution.**

***Keywords—Turner Syndrome; probabilistic networks; classification techniques based in decision trees***

## I. INTRODUCTION

This study includes processing knowledge concerning data storage and manipulation by the machine so as to be used for troubleshooting, specifically, as an adjuvant in cytogenetic diagnosis of genetic diseases. Furthermore, it aims to apply the knowledge of cytogenetic and informatics to improve the quality of genetic diagnosis.

The study presents a numerical approach to treat uncertainty by the calculation of probabilities, the reasoning being based on probabilistic inferences, i.e, the calculation of the conditional probability of an event, given the available evidence and applying the Bayes Theorem.

In general there are several evidences and direct application of this numerical approach which is questionable facing the problems of complexity for big-sized applications, as it requires a huge array of conditional probabilities are estimated and provided for the system, preventing the acquisition of knowledge and implying high time requirements, storage, capacity and computing power to process information of interest [6].

It is in the outlined context above that this study fits in, proposing the use of machine learning algorithms to extract knowledge models - probabilistic networks - from databases - so this new information may help the expert in the acquisition stage of knowledge that make up the process of building a supporting system to decision making [5].

## II. PROBABILISTIC DISTRIBUTION USING BAYESIAN NETWORK

The Bayesian odds are often used in statistical inferences to specify a previous knowledge and combine this knowledge with the data available through Bayes Theorem [24]. Bayes's formula then provides a rule by updating previous probabilities, retrospectively when the data are known and analyzed. Given a set of data for evaluating the fitness, the best result (the fittest) is defined as the most likely model data, respecting previous knowledge of the problem understanding [3].

The probabilistic distributions occurring in uncertain causal relationships within a problem domain. The probabilistic reasoning runs on these relationships using the Bayes theorem, which can be expressed as follows [7]:

$$P(Xi \mid Y) = \frac{P(Xi)P(Y|Xi)}{P(X1)P(Y|X1) + P(X2)P(Y|X2) + P(X3)P(Y|X3) + ... + P(Xn)P(Y|Xn)}$$

Para $i = 1, 2, 3, ..., n$

Probabilistic inference are used in the propagation algorithms of beliefs in Bayesian networks, and can be either causal, that part from the causes for the purpose; diagnostic parting from the effects to causes; inter-cause when discriminate between the causes of a common effect or mixed, characterized by the combination of two or more types previously mentioned [15].

In probabilistic inferences, we calculate the probability of an event, given the evidence found on the network.

## III. STATE OF THE ART TECHNIQUE IN AI AND DATA MINING

The primary objectives of the research are to develop efficient inference techniques for use in information system, for which it is necessary the availability of valid knowledge model [4]. As a result, rose up techniques of data mining and methods of knowledge based on Bayesian networks technology to extract valid knowledge models from the database.

For [4], to generate knowledge Bayesian models, many algorithms found in the literature require the following information: a) the list of variables, b) the arrangement of these variables, and c) Independence and dependency relations between them, i.e., the causal relationships. With this information passed to the learning software, it is assumed that

the user has a deep knowledge of the network to be generated, an assumption that is not always true.

### A. Knowledge Discovery in Database - KDD

The terminology knowledge discovery in databases was proposed in the first KDD workshop in 1989 to emphasize that the final product of the process of discovery in database was knowledge [25]. KDD has become, then, a specific interdisciplinary field that emerged in response to the need for new approaches and solutions to enable the analysis of large and complex databases [26].

The solution in data organization can be applied in the construction of Data Warehouse (DW), which allows for storing information, previously dispersed, through the identification, understanding, integrating and aggregating of the data in order to position them in the most appropriate location, to meet the organizational strategy of the company [13]. For the extraction of knowledge from the organized data are used mining tools known as data (MD), which may incorporate statistical techniques, probabilistic inferences and / or of AI, capable of providing responses to discover new knowledge in large databases.

### B. Methodology CRISP-DM

The Daimler Chrysler , SPSS and NCR industries created the CRISP-DM consortium - Cross Industry Standard Process for Data Mining and proposed a reference model for the process of data mining [19] non-proprietary and available free of charge [16]. Figure 1 describes the steps of the CRISP-DM methodology that are highly representative and used in the world market.

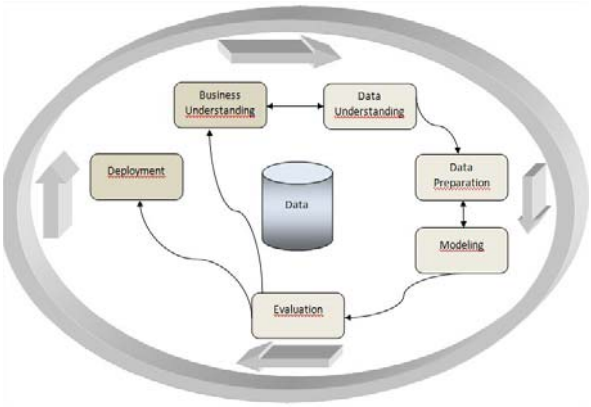


Figure 1. The process model forseen by the consortium CRISP-DM can be summarized through the lifecycle of the data mining process.

## IV. Tools and Techniques Used

### A. UnBMiner®

As [16], UnBMiner® is a visual and interactive environment, programmed in Java, platform independent, for carrying out the process of data mining. The framework automates much of the process CRISP-DM. More specifically, it covers part of the data preparation phase (pre-processing module UnBMiner®), Phase modeling (modeling formalisms based on Naïve Bayes, CNM, decision tree - algorithms ID3 and C4.5 - and multilayer neural network with backpropagation) and evaluation phase (evaluation module of UnBMiner®).

### B. UnBBayes®

As [14], the UnbBayes is a visual and interactive environment for editing and compilation of Bayesian networks (BN)

UnBBayes is developed in Java and documented with Javadoc e JavaHelp. The system is distributed freely for non-commercial use, under license GNU GPL[1]

The software UnbBayes supports the learning of the topology and the probabilities of Bayesian networks with application of search algorithms and scoring or analysis of dependencies in large databases.

### C. Classifiers based on Bayesian Networks

For [10] the network topology Naïve Bayes is represented by a tree where the root unit height is the class variable and the leaves are variable attributes.

The learning algorithm of the classifier TAN [18] first builds a structure of a tree with the variables in X \ {“TURNER”}; running test where they propagated information of mutually conditioned "TURNER". After add a link of class-node “TURNER” for each attribute-node, a structure similar to that of Naïve Bayes (i.e., the class-node is a father to all attribute-node).

The classifier BAN is an extension of the classifier TAN allowing attributes give a form the arbitrary, acyclic graph and oriented [18].

## V. Data Collection

Data collection was performed and the cases were treated in LaGene/SULEIDE/SES-GO and characterized a measurement of part of a population. The statistics calculated from the samples were used to predict various population parameters [20].

Data were collected and analyzed by the technical specialist on request of laboratory tests. Although the number of intense laboratory requests, we identified two realities that hamper data collection consistency.

The first difficulty is in cases that were indicated as "clarify the case" because the attributes were not identified what determines their exclusion.

The second difficulty was the case where the technical expert identified the attributes, thus, not formed a correlation between the attributes and their classes, in these cases there was also data deletion. Therefore, were identified 84 cases where the class is TURNER.

### A. Signs and Symptoms of Turner Syndrome

The Turner Syndrome (TS) occurs in approximately 1:2130 live births female [17; 11] and is due to the presence of one X chromosome and partial or total loss of the second sex chromosome. Despite the wide clinical polymorphism, it is

[1] General Public License (http://www.gnu.org/licenses/gpl.html)

considered that currently short stature, sexual infantilism and peripheral lymphedema are the striking clinical findings in ST. Its clinical sign most obvious and easily observed is short stature [2], which ranges on average between 142 and 146.8 cm [12; 8].

The gonadal dysgenesis is also an important signal in the ST, leading to secondary signs as primary amenorrhea, delayed pubertal development and infertility [1].They may also be subject to some congenital anomalies, including cardiovascular and renal problems, hearing loss, osteoporosis, obesity trend, and hypertelorism nipples.

Large variability of dysmorphic signs are also observed, as short neck and / or winged, broad chest and shield, cubitus valgus (from the Latin cubitus valgus), low implantation of the hair at the nape, prominent ears and low-set, fingernails hipoplástics, strabismus, epicanthal folds, among others [17; 9; 22; 21].

When there is clinical suspicion of ST, the appropriate test to confirm the diagnosis is the G-band karyotype, which allows the identification and analysis of chromosome. For karyotyping runs were cultured lymphocytes, in short-term (72h) obtained from the peripheral blood of the patient. Approximately 50% of cases, the notation classical karyotype is 45, X, which includes monosomy X , as shown in Figure 2.

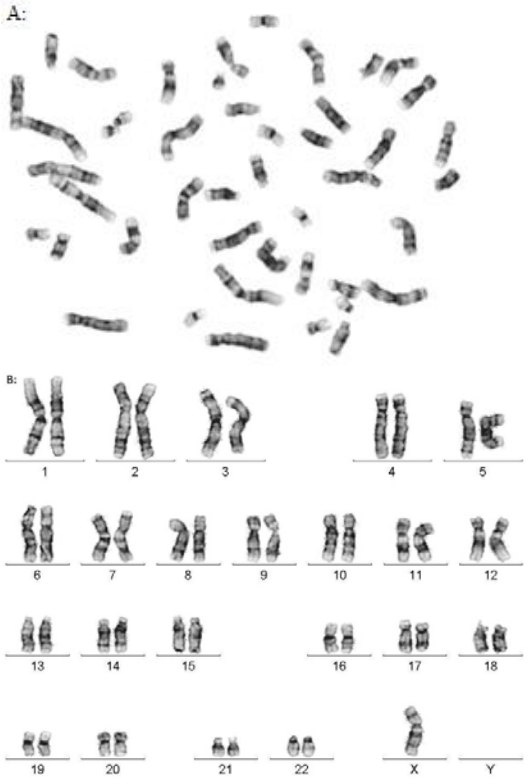


Figure 2. Human karyotype showing monosomy X, the most common chromosomal abnormality observed in Turner syndrome. A: Metaphase chromosomes containing spread, which were obtained after culturing lymphocytes for 72 hours and stained by GTG (Giemsa Trypsin +) B: chromosome pairing indicating the absence of a sex chromosome, resulting in the notation karyotype 45, X, corresponding the cytogenetic diagnosis of Turner syndrome. Source. Cytogenetics Laboratory, Human Molecular Genetics (LaGene), photograph kindly provided by Dr. Claudio C. Silva.

## VI. Result and Discussion

The proposed objective of the study was to follow the steps of the methodology CRISP-DM, therefore, it is necessary attention to the attributes for the scope of our study. Based on data collected by the technical expert, the attributes were identified (signs / symptoms) and class (syndrome) of the problem.

Using UnBMiner® was possible to perform the steps defined by the CRISP-DM. The first step was made by preprocessing, where the attributes were identified.

Both the class and attributes received values "YES" or "NO" and converted to "1" and "0" respectively. We used a class: "TURNER". The class TURNER was diagnosed by 9 independent attributes, represented by: "female, short stature, chest shield, gonadal dysgenesis, hypoplastic nails, cubitus valgus, winged neck, hypertelorism nipples and a tendency to obesity." There were 84 patient records for cases of ST.

It was necessary to clean the data, thus we also used a regression using the step-wise method to perform the selection of variables and identify attributes that are relevant to the class. As the number of cases was small, so it was necessary to apply methods to validate the data with consistent results, in this case the method of choice was 4-fold cross-validation [23].

There has been chosen an algorithm and a tool – UnbBayes – able to propagate evidence and develop efficient inference techniques able to originate appropriate techniques to combine the expert knowledge with defined data in a databank.

## VII. Conclusions

The present study demonstrated that the theory involving the Bayesian network has provided consistent results that allow the construction of knowledge-based systems. Thus, it was possible to integrate learning and propagation of evidence, yielding good results.

The problem encountered in building the Bayesian network and the forming of an expert probabilistic system was to obtain knowledge, in that, most data which served as a parameter for obtaining the results contained incomplete information, making the number of cases useful for the experiment were reduced.

A conclusion about the best solution to work out the shown problem in this study that was the Naïve Bayes model, because this one presented the greatest accuracy. The decision - ID3, TAN e BAN tree models presented solutions to the indicated problem, but those were not as much satisfactory as the Naïve Bayes. However, the neural network did not promote a satisfactory solution.

## References

[1] AM. Pasquino, F. Passeri, I. Pucarelli, M. Segni, G. Municchi, Spontaneous pubertal development in Turner's syndrome. Italian Study Group for Turner's Syndrome. J Clin Endocrinol Metab. p. 1810–1813, 1997.

[2] BM. Lippe, Turner Syndrome. In: Sperling MA, ed. Pediatric Endocrinology. Philadephia:WB Saunders Company. p. 387-421, 1996.

[3] B-T. Zhang, A Bayesian framework for evolutionary computation. Proceedings of the 1999 Congress on Evolutionary Computation (CEC99), 1:722-728; 1999.

[4] C. Koehler, MR. Vicari, CD. Flores, MS. Nassar, Mineração de Rede bayesianas a partir de Base de Dados Médicos: Proposta de Algoritmo. Univ. de Caxias do Sul (UCS), Caxias do Sul. Brasil; 2004.

[5] C. Koehler, SM. Nassar, Modelagem de Redes bayesianas a partir de Dados Médicas. Symposio de Informática y Salud; 2002.

[6] CD. Flores, Fundamentos dos Sistemas Especialistas. Porto Alegre, Rio Grande do Sul, 2002.

[7] E. Rajabally, P. Sen, S. Whittle, J. Dalton, Aids to Bayesian Belief Network Construction Second IEEE Internattonal Conference On Intelligent Systems; June 2004. p. 7803-8278.

[8] GG. Massa, M. Vanderschueren-Lodeweyckx, Age and height at diagnosis in Turner syndrome: influence of paternal height. Pediatrics, p. 1148-52, 1991.

[9] J. Batch, Turner syndrome in childhood and adolescence. Best Pract Res Clin Endocrinol Metab. p. 465-82, 2002.

[10] J. Cheng, R. Greiner, Comparing Bayesian Network Classifiers, Proceedings of the fifteenth international conference on uncertainty in artificial intelligence, 1999.

[11] J. Nielsen, M. Wohlert, Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. Hum Genet. p. 81–83, 1991.

[12] JG. Hall, DM. Gilchrist, Turner syndrome and its variants. Pediatr Clin North Am. P.1421-44, 1990.

[13] M. H. Brackett, The data warehouse challenge: taming data chaos. New York: John Wiley & Sons, 1996.

[14] M. Ladeira, DC. da Silva, FJF. Lima Jr., MS. Onishi, RN. Carvalho, WT. da Silva, Ferramenta Aberta e Independente de Plataforma para Redes Probabilísticas; 2002. M. Ladeira, Diagrama de Influências Múltiplo Secionado. [Tese]. Porto Alegre:[s.n], 2000.

[15] M. Ladeira, MHP. Vieira, H.A Prado, RM Noivo, DBS. Castanheira, UnBMiner® – Ferramenta Aberta para Mineração de Dados, Univ. de Brasília; 2005.

[16] MVN. Lipay, B. Bianco, ITN Verreschi, Gonadal dysgenesis and tumors: genetic and clinical features. Arq Bras Endocrinol Metab. vol. 49, nº 1. p. 60-70, 2005.

[17] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian Networks Classifier. Machine Learning, 29, p. 131-161; 1997.

[18] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, CRISP-DM 1.0: Step-by-step data mining guide. SPSS, 1999.

[19] R. Larson, B. Farber, Estatística aplicada; trad. Cyro de Carvalho Patarra – São Paulo: Prentice Hall, 2004.

[20] [21] RG, Rosenfeld, Turner syndrome: a guide for physicians. California:The Turner Syndrome Society; 1992.

[22] RN. Rieser, LE. Underwood, Turner Syndrome: a guide for families. California:The Turner Syndrome Society; 1992.

[23] RR. Bouckaert, E. Frank. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. Ed. Springer Berlin/Heidelberg. vol.3056/2004, p.3-12, 2004.

[24] SJ. Press, Bayesian Statistics: Principles, Models, and Applications, Wiley; 1989.

[25] UM. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework. Second International Conference on KD & DM. Portland, Oregon; 1996.

[26] W. Romão, Descoberta de Conhecimento Relevante em Banco e Dados sobre Ciência e Tecnologia. [Tese]. Univ. Federal de Santa Catarina, Santa Catarina-Brasil; 2002.

## AUTHOR PROFILE

Hugo Pereira Leite Filho received the master degree in Environmental Health Sciences, with emphasis in AI and Bayesians Networks in 2006. He is actually teaching in the institute of Information System in University State of Goiás – Brazil.