Medical Data Mining: a case study of a Paracoccidioidomycosis Patient's Database

Eduardo Liboredo Ferreira
and Herbert Rausch
and Sergio Campos
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais,
Brasil, 31270-901
Telephone: +55 (31) 3409-5566
Email: eduardoferreira@ufmg.br,
scampos@dcc.ufmg.br

Alessandra Faria-Campos
INMETRO
Xérem-Duque de Caxias, Rio de Janeiro,
Brasil, 25250-020

Enio Pietra
and Lilian da Silva Santos
Medical School
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais,
Brasil, 30130-100
Telephone: +55 (31) 3409-9300

***Abstract*—Data mining applied to medical databases is a challenging process. The unavailability of large sources of data and data complexity are some of the difficulties encountered. This is especially true for rare and neglected diseases. Those databases are, in general, relatively small, wide and sparse, making them very challenging to analyze. There are also ethical, legal and social issues regarding privacy and clinical validation of the findings. This work proposes a way of dealing with this challenge with a case study of data mining applied in a Paracoccidioidomycosis (PCM) patients database. Paracoccidioidomycosis (PCM) is a typical Brazilian disease, caused by the yeast *Paracoccidioides brasiliensis*. This disease represents an important Public Health issue, due to its high incapacitating potential and the amount of premature deaths it causes if untreated. This paper discusses methods for the analysis of this complex dataset, to help increase the understanding of both the disease and this type of data. Despite the challenges of the dataset, some interesting findings were made being: flaws in form filling protocols, notably the lack of chest X-ray in 40% of the records; the discovery of a possible new relation between smoking habits and PCM evolution time. The average evolution time for smoking patients was 2.8 times longer; the successful classification/prediction of the cutaneous form of the disease with a 93% precision rate are some of the discoveries made.**

I. Introduction

Data mining applied to medical databases is a challenging process. The unavailability of large sources of data and data complexity are some of the difficulties encountered. The examination protocols are, in general, complex and have several attributes. Different tests and exams are requested based on doctors personal experience and resource availability. Patients often fail to comply with the follow up procedures leaving the medical records incomplete. This tends to produce datasets that are difficult to analyze and require the use of multiple tools and techniques to be efficiently explored. There are also ethical, legal and social issues regarding privacy and clinical validation of the findings. This is especially true for rare and neglected diseases. Data mining use is, however, increasing important in clinical and health care fields. It can help health care insurers detect fraud and abuse, health care organizations make customer relationship management decisions, physicians identify effective treatments and best practices [7] and help researchers identify important disease traits for diagnose and treatment [2]. This work presents a case study of a Paracoccidioidomycosis database. This is one of the largest case studies for the disease in the world, the patient data corresponds to 35 years of data collection in the maximum endemicity region of Brazil.

Paracoccidioidomycosis (PCM) is a typical Brazilian disease, caused by the fungus *Paracoccidioides brasiliensis*. The most common form of infection is through inhalation of the mycelial form. It causes infection of the lung epithelium and can spread to other organs. The disease mainly affects farm workers who are exposed to contaminated soil during labor, but it is known to affect non farm workers as well [13]. This disease represents an important Public Health issue, due to its high incapacitating potential and the amount of premature deaths it causes if untreated [10]. The analysis and management of PCM related data presents several challenges. One of the challenges is related to data acquisition during patient evaluation and diagnosis. The Center of Training and Reference on Infeccious-Parasitary Diseases from the Federal University of Minas Gerais (CTR-DIP-UFMG) has developed a protocol for clinical analysis of PCM patients. This protocol includes a large number of clinical variables that are assessed in each medical examination, including x-ray and serology tests, which are also used in tracking the disease progression. Currently, there are no reliable clinical parameters for PCM to establish the treatment duration nor predict the disease relapse. The use of data mining techniques and Business Intelligence (BI), can help to find patterns and useful information otherwise invisible.

Several tools are available for data analysis. Some specialize in multi-dimensional analysis, others in data mining and others in statistical analysis. Researchers must often use different tools, under different environments and platforms (web, desktop) to obtain all the relevant data. In this work, we have used the tools Mondrian [8], WEKA [4] and R [12] to filter, select and use clustering and classification techniques on the information available to explore the correlation between its

978-1-4799-6644-8/14/$31.00 ©2014 IEEE

variables. The records used are composed of first examination data of patients diagnosed with PCM. A total of 227 patients charts with 314 items each compose the database.

II. Data Mining

Data mining can be defined as the computational process of discovering patterns in data and present them in an understandable and useful way [6]. It aims to uncover patterns and relations that are invisible by manual processing. In order to be useful, the raw data must be collected, pre-processed and stored in digital format. This section summarizes the data mining process used for the PCM database, introducing and explaining basic concepts. Details of the individual methods used are described in more details on each section following.

The collection of data, pre-process and storage is called Extraction, Transformation and Load (ETL). The ETL is one the most important steps in data mining [16], [6]. During the ETL, data from multiple sources are formated, standardized, and minor issues and human error when collecting data, such as the same value with different spellings, are corrected. Section *Database and ETL* describes the details of the process applied to the PCM dataset.

To optimize analysis, techniques for reducing the dimensionality of the database were employed. Through this process, irrelevant and redundant fields are eliminated, reducing computational costs and generally improves the quality of the analysis [6], [3]. To perform the reduction two methods have been used, *Frequency Distribution Analysis* and *Attribute Selection*.

For the analysis of frequency distribution, the tool Mondrian was used. Mondrian is an open source On line Analytical Processing server (OLAP) [8]. It is a versatile tool, offering a simple and intuitive interface and allows the user to easily visualize and navigate through the data. The frequency distribution shows the most common value of an attribute. It is important to retrieve this information because fields that have a highly dominant value tend to be irrelevant for clustering and classification, often introducing noise in the analysis [5]. The details are discussed in the subsection *Reduction of Dimensionality*.

The next step for dimensionality reduction was Attribute or Feature Selection. It is a process for finding the best subset of relevant features related to a class(attribute) for model construction. The models used in this work were clustering and classification trees, discussed later on. Two different approaches have been taken for this process . The first method was unassisted, with a progressive selection based on probability [9], to select the most relevant attributes on the dataset based on a statistical score. The results and detail are discussed in the respective subsection *Attribute Selection*. The models for the unassisted analysis were built using clustering algorithm k-means.

Clustering algorithms divides the instances, patients in this case, into natural groups, presumably revealing important attributes that separates them from the other groups [16]. This technique applies when there is no class or attribute to be predicted and it can highlight the most relevant attributes [5], [4]. The details and results are described in the section *Unassisted Analysis: Clustering*.

The second method for attribute selection is an assisted method. A set of attributes, called base attributes, have been selected by specialists as clinically relevant and used to further refine the dataset for analysis. Details on methodology and results are discussed in the section *Expert Assisted Analysis* and following subsection. For the assisted analysis, the construction of classification models has been performed, more specifically, classification trees were built.

Classification trees are a set of rules and steps that leads to the prediction of the possible value of a target attribute [16], [11]. A graphical representation of a decision tree is shown in III. The fundamental difference between these methods – clustering and classification trees — is that clustering is an undirected method for grouping and classification while the decision tree is directed, predicting the value of the intended, or target, attribute.

III. Database and ETL

One of the most important steps in data mining is the process called Extraction, Transformation and Load (ETL). During the ETL, the data is gathered, filtered and formated for the intended analytical tools. The raw database for this study contains the clinical data from 227 PCM patients made available by CTR-DIP-UFMG in SSPS format. The database consists of 314 attributes per patient, with 09 numeric type attributes (age, time of progression, RCD size, AX size, boyd size, lesion size, lesion area, inactivation time, treatment duration) and the remaining attributes are nominal.

Two databases have been prepared, one for the On Line Analytical Processing (OLAP) analysis and the second for data mining techniques, which includes attribute selection, clustering and building classification trees. The ETL process of the raw data has followed these steps: 1. Conversion of the SSPS files into CSV files; 2. Elimination of irrelevant fields to the analysis, such as patient personal information and protocol number; 3. Standardization of spelling in applicable fields. 4. Uploading the CSV into a MySQL table. This processes left the database with 301 attributes.

A. Reduction of Dimensionality

The Reduction of the dimensionality of the data by deleting unsuitable attributes improves the performance of learning algorithms and, more important, yields a more compact and easily interpretable representation of the target concept, focusing users attention on the most relevant variables [16], [3]. To reduce the dataset, elimination of redundant or irrelevant features have been conducted. Redundant features are those which provide no more information than the currently selected features and irrelevant features are those that provide no useful information in any context. The methods used to reduce dimensionality were *Frequency Distribution Analysis* and *Attribute Selection* [6], [3], discussed in more details in following sessions.

1) Frequency Distribution: The first step in reducing dimensionality was a frequency distribution analysis of the data, using the OLAP tool Mondrian [8]. The distribution of frequency analysis shows how many times a value is present in the records and gives a general picture of data distribution. Fields with a high concentration of a single value tend to be

TABLE I. FREQUENCY DISTRIBUTION: ATTRIBUTES, RESPECTIVE MOST COMMON VALUE AND IS FREQUENCY.

Attribute	Value	frequency %
PCM renal	normal	95.2
PCM pancreatic	normal	95.2
PCM splenic	normal	93
PCM genital	normal	94.7
PCM adrenal	normal	93.9
PCM intestinal	normal	93.9
PCM bone	normal	93
PCM gastric	normal	93.1
PCM liver	normal	92.1
PCM limphatic	normal	91.3
PCM neurologic	normal	90.4
positive Serological test	Unfilled	98.69
Boyd size (cm)	Unfilled	85.15
lesion area (cm^2)	Unfilled	78.6
facial X-ray	Unevaluated	93.5
altered Lactate	Unevaluated	92.1
altered CPK	Unevaluated	91.7
altered CK-MB	Unevaluated	90.8
positive HBeAg	Unevaluated	90.8
positive Anti-HBc	Unevaluated	90.5
altered Lipase	Unevaluated	88.2
positive CMV	Unevaluated	87.8
positive Anti-HIV (Elisa)	Unevaluated	87.3
positive Chagas	Unevaluated	87.3
altered Uric acid	Unevaluated	86.5
positive Anti-HAV	Unevaluated	86.34
altered Culture	Unevaluated	86.4
altered Spirometry	Unevaluated	86.5
positive Anti-HCV	Unevaluated	86.3
altered Amylase	Unevaluated	85.6
Anti-HBs positiv	Unevaluated	85.6

irrelevant or even detrimental to mining algorithms as they create noise [16]. Fields with more than 70% of missing values have been discarded. Different frequency ratios have been tested for elimination and the 70% threshold obtained the most relevant results. With a dataset as complex as the PCM, a careful approach while discarding fields had to be taken. Approximately 29% of the database used is composed of missing values. This accounts for 19422 of 66822 total fields. A total of 109 attributes (36% of 301 used) have missing values count above 30%. Setting the threshold too high for missing values would incur the risk of eliminating too many relevant fields.

An interesting observation is that of the 17 types of PCM tested in the examination protocol, 11 types are not significantly observed in the dataset as shown in table I. The table also shows other attributes with high concentration of a single value. From the 314 initial attributes, 52 were discarded, 262 remained for analysis. Only fields with missing values were discarded from the dataset. All 17 types of PCM were maintained.

2) Attribute Selection: Attribute selection or feature selection is a process for finding the best subset of relevant features related to a class(attribute) of the model. The feature selection algorithm selects a group of features that are closely related to a selected class and contains relevant information. Two different approaches have been taken for this process. The initial attribute selection was conducted with a progressive selection based on probability [9], to select the most relevant attributes on the dataset based on a statistical score. It is an unassisted method, one of the reasons why it has been chosen as the starting point. The second approach consisted in selecting attributes based on specialists opinions on which attributes were clinically relevant. This method is discussed in the session *Expert Assisted Analysis*.

TABLE II. CLUSTERING OF RELEVANT FIELDS FROM DATA SUBSET "LESION MUCOSA". FIELDS MARKED WITH "-" ARE UNEVALUATED, FIELDS MARKED WITH "Y" ARE ALTERED TESTS OR PRESENT SYMPTOM AND MARKED WITH "N" ARE ABSENT OR NORMAL.

Attribute	#0	#1	#2	#3	#4
	41(18%)	67(30%)	19(8%)	47(21%)	53(23%)
evolution time	11.8537	10.6567	17.7368	36.8298	5.2264
skin lesion	Y	N	Y	N	N
smoking	Y	Y	-	Y	N
pcm mucosa	Y	Y	Y	N	N
mucosa lesion	Y	Y	Y	N	N

Based on the initial unassisted selection, four attributes have been selected as the most relevant: "Gender", "Vomiting", "Skin Lesion" and "Mucosa lesion". Further attribute selection has been conducted using this four attributes as base to create a subset for cluster analysis. The data subset consists of the attribute used as base for the selection and the attributes given by the selection algorithm. For example, the complete subset for the "Mucosa lesion" is : "mucosa lesion" "evolution time", "skin lesion", "nasal obstruction", "difficulty swallowing", "smoking habits", "blood pressure while standing", "mouth lesion" and "pcm mucosa". This subset was used for clustering as shown in table II. The algorithm used for attribute selection was the Correlation-Based Feature Selection (CfsSubsetEval) [1] implemented on Weka. From this point on, all relevant findings have been sent to specialists for further analysis.

IV. ANALYSIS

A. Unassisted Analysis: Clustering

Clustering techniques apply when there is no class to be predicted but the instances are to be divided into natural groups [16], [4], [5]. In this work each subset of selected attributes was clustered using k-means, available in Weka. K-means is a well know algorithm that requires a definition of the number of clusters (K value) *a priori*. The initial value of K was obtained using the Expectation Maximization (EM) algorithm, which can give a good estimation of the optimal number of clusters [16], [6]. Further filtering was made in the initial clusters, to identify and remove irrelevant fields and optimize the k value. Some clusters showed correlation between attributes, which have been submitted to specialists for analysis. The clusters for the subsets "Gender", "Vomiting" and "Skin lesion" showed no relevant results. Other algorithms were used for clustering the subsets, DBScan and hierarchical clustering, but showed no relevant results.

Table II presents the most relevant correlation in the initial analysis. It shows a reduced scenario, considering only the five relevant fields, which are evolution time, skin lesion, previous smoking habits, PCM mucosa and mucosa lesion. It shows the difference in disease evolution time between smoking and non-smoking patients. The optimal number of clusters found (k value was 5).

The average evolution time for non-smoking patients(clusters 2 and 4) was 7.04 months and smoking patients got an average time of 19.83 months(clusters 0, 1 and 3), 2.81 times longer. Analyzing patients with PCM mucosa the relation is 7.36 months for non-smoking and 12.55 for smoking. There

is a described relation between PCM and smoking in [14], [15] but not directly to its progression. The field ”evolution time” represents for how long the patient contracts the disease and it continues to progress. As described before, the most common infection method for PCM is the inhaling of contaminated soil. It starts mostly as the pulmonary version of the disease. If untreated, it can spread to other organs, like liver, limphonodes and even the nervous system. This study suggests that smoking habits helps the disease progression. Further investigation is needed to determine the nature of the influence.

Although no other subset produced relevant results, one of the clusters of the analysis had, as one of the main traits, the lack of X-ray test results. In total, 40% of the patients had ”unevaluated“ for the chest X-ray parameter. This realization came as a surprise to specialists since the chest X-ray is one of the primary tests in examination protocol.

V. Expert Assisted Analysis

After the initial analysis, a directed approach was taken. Of the 262 attributes left after the frequency analysis, 23 have been chosen by the specialist staff as relevant attributes for further correlation investigations. Gender, age, disease relapse, rural contact, smoking habits, drinking habits and the 17 forms of the disease have been indicated. A similar method to the previous analysis has been performed, consisting of attribute selection using the indicated attributes as base and discarding the attributes considered irrelevant. A total of 112 attributes, each showed significant correlation to at least one of the 23 base attributes, remained for the analysis. The filtered database was then used to build decision or classification trees. It is worth to note that other approaches for classification were tested. All attributes were also classified using bayesian networks, notably the NaiveBayes and BayesNet algorithms implemented in Weka. The results for the trees were largely superior and are described in the following session.

A. Classification trees

The decision tree is a data mining technology suitable for performing classification and prediction. The decision tree can produce results according to different variables by repetition that can thus be used to analyze the characteristics, similarities and differences in data [11]. As mentioned before, classification trees are a directed approach used for the prediction of the possible value of a target attribute [16], [11] In this work, Weka’s integrated j48 tree algorithm has been used for the classifiers. It is based on the widely used C4.5 algorithm (for details, reffer to [11]). The j48 algorithm was chosen for its versatile characteristics. It classifies nominal attributes and handles numeric attributes as wells as missing values, avoiding the need to discretize numeric fields and impute data. The trees were built with the 10 fold cross-validation option, which consists in dividing the dataset in 10 parts, using 9 parts for learning and 1 part for testing, then rotating and using a different part for testing and a different set of 9 parts for learning, until every one of the 10 parts has been used for testing. The cross-validation method gives a better estimation of real world use of the rules set for building the tree. Using the whole data set for learning may give an overly optimistic performance indicator [16]. Figure 1 shows the graphical representation of the pruned classification tree used for the attribute PCM skin (cutaneous form of the disease). Pruned trees are simplified versions of the classification tree. They discard the branches that are less relevant, with less predictive power, and tend to reduce the noise of the analysis. They also provide a more readable and friendly visualization of the rules for classification. For comparison, the tree in the figure 1 has 4 leaves and total size of 7. The unpruned version of the same tree has 19 leaves and a total size of 31.

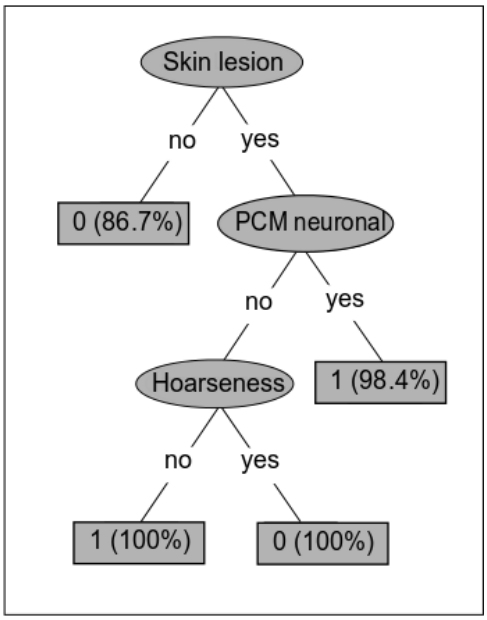

Fig. 1. Graphical representation of a pruned classification tree, built for the attribute PCM skin. The numbers that appear in the rectangles are the probable classification for the attribute PCM skin and its respective probability. 0 is negative for PCM skin and 1 is positive.

B. Classification Results

1) PCM skin: Table III shows the results for the pruned tree j48 tree, classifying the attribute PCM Skin(cutaneous). The first part shows the total number of correctly and incorrectly classified instances. The second part of the table displays more detailed information on the real quality of the classifier. In this case, it shows a 0.932 of true positives (tp rate) for class 0. This means that the instance was correctly classified as 0 in 93.2% of the tests. The false positives (fp rate) for class 0 is 0.072. This means that in 7.2% of the classification attempts, the instance was incorrectly classified as 0. The ideal scenario is a rate of 1 for tp and 0 for fp. The third part of the table is the *confusion matrix*. The matrix displays the number of correctly classified instances by class. The line for class ”negative” shows that 65 instances were correctly classified as negative for cutaneous PCM, and 5 were incorrectly classified as positive for cutaneous PCM. The line for class ”positive” shows that 128 instances were correctly classified as positive for pcm skin, and 10 were incorrectly classified as negative for cutaneous PCM. This results show that it is possible to predict the cutaneous form of PCM with high accuracy making it possible, for example, to build decision support system for this attribute, helping in the diagnosis and treatment of the disease.

TABLE III. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE PCM SKIN ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP = RATIO OF TRUE POSITIVES; FP = RATIO OF FALSE POSITIVES.

10 fold cross-validation for attribute PCM skin		
correctly classified attributes	196	92.9%
incorrectly classified attributes	15	7.1%

tp rate	fp rate	precision	class
0.932	0.072	0.872	0
0.928	0.068	0.962	1
0.929	0.07	0.931	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
65	5	negative
128	10	positive

2) Gender: Gender is an important attribute for PCM attribute correlation. The impact of female hormones in the progress of the disease, providing protection to women in reproductive age has been described in the literature [9], [15], [14]. Table IV shows the results for the pruned tree j48 tree, classifying the attribute *Gender*. The overall accuracy for gender is approximately 82%. But the accuracy for individual classes, 0 (males) and 1 (females), are very different. The precision for correctly classifying gender as 0 is 0.86 and the precision for classifying gender as 1 falls to 0.47 and the true positive rate to only 0.263. One of the factors that contributes to this variance is the composition of the dataset: 177 males and 38 females. In this case, using the full dataset as training improved the overall accuracy to almost 87% and the precision for classifying instances as 0(male) to 0.92 and 1(female) jumped to 0.625. The attributes selected by the classification algorithm for Gender classification rules are: "PCM mucosa", "PCM lungs" (pulmonary), "PCM limphonodes", "PCM skin", "swollen limphonodes" and "aortic systolic murmur". These results explicits the relation between disease type and manifestation and gender described in literature. It is important to note that both trees, 10-fold cross validation and full training set, used the same rules for classification.

TABLE IV. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE GENDER ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP IS THE RATIO OF TRUE POSITIVES, FP FALSE POSITIVES.

10 fold cross-validation for attribute Gender		
correctly classified attributes	176	81.9%
incorrectly classified attributes	39	18.1%

tp rate	fp rate	precision	class
0.938	0.737	0.856	0
0.263	0.062	0.476	1
0.819	0.618	0.789	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
166	11	male
10	28	female

3) Relapsing PCM: As previously discussed, there are no current clinical parameters to help predict the possible relapse of PCM. This is another attribute of clinical interest that showed promising results. Table V shows the correct classification was, overall, at 73.8%. The true positives for relapsing PCM (instance classified as 1) show a ratio of 0.932 and the overall precision at 0.71.

The attributes used as rules for classifying the relapsing disease are "intestinal PCM", "global leukocyte count", "treatment time", "treatment with amphotericin b", "chest X-ray" and "disseminated PCM". These attributes are important clinical features linked with disease progression and treatment, encouraging and possibly guiding further research for clinical validation. Since the data collection is still in progress, future studies on the subject may validate the parameters to safely predict a relapsing PCM.

TABLE V. PRUNED TREE USING THE J48 ALGORITHM FOR CLASSIFYING THE RECURRENCE OF PCM ATTRIBUTE. THE CONFUSION MATRIX EXPLICITS THE INCORRECTLY CLASSIFIED INSTANCES AND ITS DISTRIBUTION. TP = RATIO OF TRUE POSITIVES, FP = RATIO OF FALSE POSITIVES.

10 fold cross-validation for attribute Relapsing		
correctly classified attributes	152	73.8%
incorrectly classified attributes	54	26.2%

tp rate	fp rate	precision	class
0.241	0.068	0.583	0
0.932	0.759	0.758	1
0.738	0.564	0.709	Weighted avg.

confusion Matrix		
correct	incorrect	classified as
14	44	negative
138	10	positive

VI. CONCLUSION

Despite the difficulties in working with a complex, relatively small and sparse dataset, the results obtained by our analysis are promising. Both approaches, assisted and unassisted, revealed useful information. The discovery of a possible relation between smoking habits and disease progression encourages further research. The average evolution time for the disease progression was 2.8 times higher for smoking patients. During the clustering analysis, the lack of chest X-ray test appeared as a grouping attribute. This is a major test and 40% of the patients in the database do not have results registered. This information surprised and warned the specialists about errors in protocols and form filling procedures.

The successful classification of cutaneous PCM (PCM skin) with a high precision, 92.9%, and low false positive rates, 7%, demonstrate that data mining can be successfully applied on database with this characteristics.

The differences in the disease manifestation regarding gender described in the literature were also observed in this work. The main attributes for gender differentiation were "PCM mucosa", "PCM lungs" (pulmonary), "PCM limphonodes", "PCM skin", "swollen limphonodes" and "aortic systolic murmur". This attributes explicit the relation between disease type and gender described in literature [9], [15], [14] The model for classification of the relapsing PCM showed a precision above 70%. The attributes linked with relapsing are: "intestinal PCM", "global leukocyte count", "treatment time", "treatment with amphotericin b", "chest X-ray" and "disseminated PCM". These are important clinical features related to disease progression and treatment. These findings may help to guide further research, helping to uncover and validate the parameters for relapsing prediction. Data collection for PCM is still in progress. Despite the difficulties and complexity of the database, relevant information was obtained and we hope to encourage future research on the subject.

ACKNOWLEDGMENT

The authors would like to thank the medical staff of the Hospital das Clínicas - UFMG - for the cooperation and attention during this study.

REFERENCES

[1] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int',l Conf. Machine Learning, pp. 359-366, 2000.

[2] Breault, J.L, Goodall, C.R. & Fos, P.J. (2002). Data mining a diabetic data warehouse. Artificial Intelligence in Medicine, 26(1), 37-54.

[3] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157–1182.

[4] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: An update. SIGKDD Explorations, 11.

[5] Han, J. & Kamber, M. (2001). Data mining: concepts and techniques. Morgan Kauf-mann Publishers, second edition

[6] Hastie, T.; Tibshirani, R. & Friedman, J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction Springer, second edition

[7] Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of Healthcare Information ManagementVol 19.2 (2011): 65.

[8] Mondrian (2012). Mondrian. http://mondrian.pentaho.com/.

[9] de Moura, A. C. L. (2008). Estudo Clinico e Imunologico de Controle de Cura de Paracoccidioidomicose Cronica. PhD thesis, Universidade Federal de Minas Gerais.

[10] Portal da Saude - Ministerio da Saude - Governo Federal - Brazil http://portalsaude.saude.gov.br

[11] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[12] R (2012). R project. http://www.r-project.org/

[13] Restrepo, A.; McEwen, J. & Castaneda, E. (2001). The habitat of Paracoccidioides brasiliensis: how far from solving the riddle? Med. Mycol., 39:233241.

[14] Santos, W. A. d.; Silva, B. M. d.; Passos, E. D.; Zandonade, E. & Falqueto, A. (2003). Associação entre tabagismo e paracoccidioidomicose: um estudo de caso-controle no estado do espírito santo, brasil. Cadernos de Saúde de Pública, 19:245 253.

[15] Shikanai-Yasuda, M. A.; Telles Filho, F. d. Q.; Mendes, R. P.; Colombo, A. L. & Moretti, M. L. (2006). Consenso em paracoccidioidomicose. Revista da Sociedade Brasileira de Medicina Tropical, 39:297 310.

[16] Witten.I; Frank.E & Hall.M (2011). Practical Machine Learning Tools and Techniques.Morgan Kaufmann Publishers, third edition