DATA MINING APPLICATION TO OBTAIN PROFILES OF PATIENTS WITH NEPHROLITHIASIS

Luis Zárate, Paulo Alvarenga, Romero Paoliello, Thiago Ribeiro
Applied Computational Intelligence Laboratory (LICAP)
Pontifical Catholic University of Minas Gerais (PUC)
Av. Dom José Gaspar, 500, Coração Eucarístico
Belo Horizonte, MG, Brasil, 30535-610

Keywords: KDD, Data Mining, Discriminant Rules, Clinical Databases, Nephrolithiasis

Abstract: Nephrolithiasis is a disease that is unknown yet a clinical treatment that determines its cure. In the adult population is esteemed an incidence around 5 to 12%, being a little lesser in the pediatric band. The renal colic, caused by nephrolithiasis, is the main disease symptom in the adults and it is observed in 14% of the pediatric patients. The disease symptoms in the pediatric patient don't follow a pattern, and this makes difficult the disease diagnosis. The main objective of this work is discovery the patters of the disease symptoms and identifies the population apt to acquire it. With this objective, is applied KDD methodology determining discriminant rules for the patterns of the symptoms, and with this, select the groups of patients with those sets of symptoms. Finally, the results and the conclusions of the work are presented.

1 INTRODUCTION

Nephrolithiasis is a disease that is unknown yet a clinical treatment that determines its cure. In the adult population is esteemed an incidence around 5 to 12%, being a little lesser in the pediatric band. The renal colic, caused by nephrolithiasis, is the main disease symptom in the adults and it is observed in 14% of the pediatric patients. The disease symptoms in the pediatric patient don't follow a pattern, and this difficult the disease diagnosis. The main objective of this work is discovery the patterns of the disease symptoms and identifies the apt population to acquire it.

The diagnosis of the disease is accomplished through clinical exams and by the observation of the symptoms. Between the main symptoms we can cite: abdominal pain with not specific localization, hipertension, fast and gasping breath, nauseas, vomits, anorexy, indisposition, pulse and arterial pressure with alteration, and hematuria macro or microscopic. On the other hand, the diagnosis that seems easy due to the intrinsic symptoms characteristics, as vomit, abdominal pain, gasping breath, among others, can take to a false diagnosis. This phenomenon is understandable if we remember that those symptoms are the same of other diseases, as for example, appendicitis, acute pancreatitis, etc. Therefore, exams of laboratory are necessary to confirm the diagnosis. In this work, is discussed a strategy to discovery the disease patterns on symptoms proceeding from exams of laboratory, genetic inheritance, and by the patient's clinical observation.

The discovered patterns consider the presence or not of the symptom in the patient and they will be expressed through discriminant rules that can assist in the disease diagnosis. On the other hand, it is also important to know the groups profile of risk patients, what can be obtained through techniques of clustering, discrimination or classification. In this work will be applied the KDD methodology based on discriminant rules for discovery patters of disease symptoms and for the identification of apt groups for acquire it. The KDD Methodology (Fayyad, et. al. 1996 and Pyle, 1999) is a long procedure of specific stages, having as only objective, the discovery of knowledge in database. The KDD process involves several stages, such as:

Zárate L., Alvarenga P., Paoliello R. and Ribeiro T. (2004).
DATA MINING APPLICATION TO OBTAIN PROFILES OF PATIENTS WITH NEPHROLITHIASIS.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 104-109
DOI: 10.5220/0002632501040109

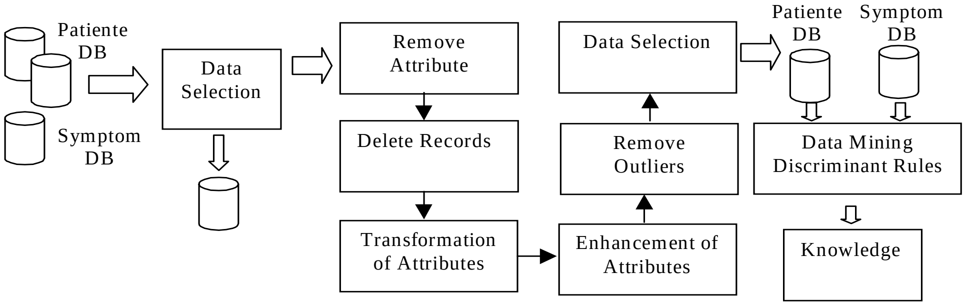

Figure 1: Applied KDD methodology

selection, enrichment, improvement, cleaning, transformation, data mining and interpretation. The application of KDD in the medical area has been treated in several works, as can be seen in Bojarezuk, Lopes & Freitas, 2001 and Werner & Fogarty, 2001.

In this paper the diferents stages of KDD will be applied in a nephrolithiasis database to obtain discriminant rules that match with the disease pattern in the patients. Those rules will make possible verify the importance of the pattern, and the symptoms description with the respective occurrence. It is also objective of this work group patients with the same profile to make possible the identification of apt patients to develop the disease through the discriminant rules.

This work is divided in 5 sections. In section 2, the nephrolithiasis database, considered in this work, is described. In section 3, the KDD process is applied, and in the last sections the results and the conclusions of the work are presented.

2 DESCRIPTION OF THE CONSIDERED DATABASES

The space problem definition to be investigated mentions the discovery patterns of disease symptoms and the identification of apt groups to acquire it. The database considered is composed by patient’s data and diagnostic that had presented or not the disease. This database presented 69 attributes with the total of 363 records. The typical record of the original database, with categorical names, is express in (1).

$$\text{Record} = \{DP, CC, SL, SC, DG, OC, OL\} \quad (1)$$

where: DP = personal data (23 attributes); CC = clinical conduct (20 attributes); SL = symptoms resulted by laboratory exams (12 attributes); SC = clinical symptoms (1 attribute); DG = genetic inheritance (4 attributes); OC = clinical observations (7 attributes); OL = laboritorial observations (2 attributes). It is evident that nor all the data are important for to the proposed analysis.

The selection of the important attributes to the problem is a difficult task (Jones, 1998). Therefore, each attribute should be analyzed in an individual way and together. For example, the attributes (CC) correspond to judgements of subjective character and therefore they were eliminated of the group of attributes. The original database needs to pass for a selection process, enrichment, enhance, cleaning and transformation, before they can be submitted to the algorithms of data mining. In the next section are described the data preparation stage and the data mining techniques as discriminant rules.

3 APPLICATION OF THE KDD PROCESS

The KDD methodologies vary from author to author but all deal with the same necessary stages for accomplishment of the process. In general lines the stages of KDD involves: Selection, Pre-processing, Data Mining, Interpretation and Visualization. Each mentioned stage involves other steps that depend on the specific problem. In the continuation will be described the stages applied to the nephrolithiasis database. The Figure 1 shows the structure of the applied stages in this work.

3.1 PRE-PROCESSING DATA SETS

3.1.1 Removing Attribute

The criteria adopted for removal of attributes were based on the database analysis and on the problem in subject:

a) The first criterion determines as irrelevant the attributes that had presented less than 3 occurrences in the total of 363 records. 18 were the total of excluded attributes.

b) The second criterion was the irrelevance of attributes that do not contribute in the space problem definition. Some of these were: related attributes to the medical conduct and the field of names. The field <medical conduct> is related with the decisions taken for the doctors in intention to treat the disease, and therefore it does not possess direct relation with the determination of its cause, and the field <name> does not have any relation with the intended analysis. 12 were the total of excluded attributes.

3.1.2 Deleting Records

The criteria adopted for exclusion of records of the database had been:

a) Elimination of the records of patients who had presented empty fields of symptoms, due to the main objective, that is discovery the disease symptoms patterns.

b) Elimination of inconsistent records. As example, the weight of a child of 300 Kg can be cited (certainly typing error) for which was opted to not correcting (established in similar records) in way to not polarize the discovered knowledge. 12 were the total of excluded records.

3.1.3 Enhancement

From the literal attribute <current clinical situation> contained in the category {OC}, had been created 5 new attributes. They were: Abdominal pain, Hematuria, Renal Colic, Asymptomatic, ITU, These fields had been created due to high frequency of occurrences.

3.1.4 Transformation of Attributes

The following transformations had been applied to some attributes of the database:

a) For the <date> fields that possessed the format: "dd/mm/yy", had been transformed for the age into years "yy". This transformation was realized in the following fields: <DateOfBirth>, <DateOfDiseaseBeggining> and <DateOfDiseaseBeginningOfControl>.

b) Fields with the measure in meters <height>, had been transformed into centimeters.

c) Attributes referring to the genetic inheritance from the father, the mother or other familiar had been substituted by an only attribute <FamiliarHereditarySucession>.

d) Attributes referring to the symptoms Hematuria (X), Macroscopic Hematuria (Y) (physical result of the symptom Hematuria) and Microscopic Hematuria (Z) (resulted of laboratory exams of the symptom Hematuria) had been grouped in an only attribute Hematuria (W) obeying the following rule:

$$w = x \vee \mathrm{y} \vee \mathrm{z}$$

3.1.5 Removing Outliers

After realized the data elimination and the data improvement, 351 were the total of valid records in the database. In way to evaluate the dispersion of the referring data to the patients' symptoms (records), the curve of Gauss was used to eliminate records that presented accented deviation in relation to the others. The adopted procedure is described as follow:

The number of realized comparisons was 61425 for the 351 records. For each combination of 2 records, the numbers of occurrences of equal symptoms for the involved patients had been counted. The expressions (2) and (3) show the adopted strategy.

$$\begin{gathered} S^{i} = \{s_1^i, s_2^i, s_3^i, \ldots, s_N^i\} \\ \text{where} \quad \mathrm{s}_{\mathrm{n}}^{\mathrm{i}} = \{0,1\} \quad \text{for} \quad i = 1, \ldots, \mathrm{M} \end{gathered} \tag{2}$$

$$\begin{gathered} \forall C^{kl} \quad \text{for} \quad k, l = 1, \ldots, M \quad \text{with} \quad k < l \\ cont^{kl} = S_j^k (notXOR) S_j^l + cont^{kl} \\ \text{where} \quad j = 1, \ldots, N \end{gathered} \tag{3}$$

where: *M* is the number of patients, *N* is the number of symptoms, S^i is the vector of symptoms of the "*i*" patient, s_j^i is the "*j*" symptom of the "*i*" patient and C^{kl} is the combination of the "*k*" patient with "*l*" patient. In this work, *M*=351 and *N*=14.

For elimination of a register the follow criteria was adopted: Records of patients with less number of occurrences than $X - 2S(x)$ (where X is the average of occurrences $cont^{kl}$ and $S(x)$ is the standard deviation of the sample) can be inconsiderate. We verified that none records were removed with this consideration due to the little data dispersion.

3.2 DATA MINING

In this section the objective is to obtain discriminant rules that characterize the symptoms patterns. Through these patterns is possible to discovery patients' profiles apt to acquire the disease. In the continuation the followed procedures will be described.

3.2.1 Obtaining Characteristic Patterns of the Disease

In this stage the objective is to identify the disease characteristic patterns. For that the data were separated in patient and symptoms keeping its relationship (1:1). See Figure 2.

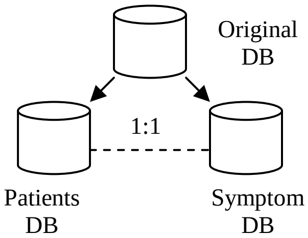

Figure 2: Databases structures

Through an analysis of the symptoms database (that contain patient's sick and not sick data), 150 distinct patterns had been observed, approximately, involving all the symptoms. In way to reduce the number of patterns, the occurrences percentage of the presence $PercS$ and the absence $PercNS$ of each symptom were calculated. The expression for calculation is given by the expression (4) and the values are shown in Table 1.

Be the vector of symptoms of the patient *"i"*: $S^i = \{s_1^i, s_2^i, s_3^i, ..., s_N^i\}$ given by the equation (2)

$$for\ \ s = 1; PercS_j = (\sum_{i=1}^{M} s_j^i)/M\ \ for\ \ \ j = 1,...,N \quad (4)$$

$$for\ \ s = 0;\ \ PercNS_j = 1.0 - PercS_j \quad (5)$$

In Table 1, *"Sick"* represents the relative frequency of sick people who present the symptom *"j"* and *"NS-Sick"* is the relative frequency of sick people who had not presented the symptom *"j"*. Through this Table is possible to obtain conclusions of the type:

$$(\forall s_j = 1)[q_j\%]\ \ with\ \ (S\_Sick)[t_j\%]\ \ \Leftrightarrow \quad (6)$$
$$(\forall s_j = 0)[q_j\%]\ \ with\ \ (NS\_Sick)[t_j\%]\forall j = 1...N$$

Table 1: Occurrences Percentage

Symptom	$PercS_j$	*S-Sick*	*NS-Sick*
s_1	0.63	0.70	0.96
s_2	0.04	1.00	0.79
s_3	0.31	0.80	0.80
s_4	0.01	0.80	0.80
s_5	0.42	0.80	0.79
s_6	0.15	0.67	0.82
s_7	0.07	1.00	0.78
s_8	0.12	1.00	0.77
s_9	0.49	0.74	0.85
s_{10}	0.20	0.87	0.78
s_{11}	0.08	0.96	0.78
s_{12}	0.05	1.00	0.79
s_{13}	0.02	1.00	0.79
s_{14}	0.49	0.68	0.91

Using the expression (6) is possible to obtain the following rules:

$$(s_1 = 1)[63\%]\ \ with\ \ (S\_Sick)[70\%]\ \ \Leftrightarrow$$
$$(s_1 = 0)[37\%]\ \ with\ \ (NS\_Sick)[96\%]$$
$$(s_5 = 1)[42\%]\ \ with\ \ (S\_Sick)[80\%]\ \ \Leftrightarrow$$
$$(s_5 = 0)[58\%]\ \ with\ \ (NS\_Sick)[79\%]$$

Notice in the Table 1, that there are small values for $PercS_j$ (for example: s_4, s_{13}). In way to reduce the number of different patterns, a criterion was applied based on a heuristic rule through which the less characteristic symptoms of the disease are retired. The rules adopted were:

$$if\ \ (0.03 < PercS_j \leq 0.10\ \ AND \quad (7)$$
$$freqSick_j < \alpha)\ \ Then\ \ s_j \notin S$$

$$if\ \ (PercS_j \leq 0.03)\ \ Then\ \ s_j \notin S \quad (8)$$

where $freqSick_j$ represents the frequency of sick people for the symptom s_j. α is a adjustment parameter that determines how much a symptom can be considered or not. In this work it was chosen as $\alpha = 0.85$. With the heuristic rule proposed, the number of symptoms reduced to 12. To obtain patterns classification, the characteristic weight was calculated for each vector S^i through the equation: (9):

$$Weigth^i = \sum_{j=1}^{N} \left\{ \begin{matrix} PercS_j & paras^i_j = 1 \\ PercNS_j & paras^i_j = 0 \end{matrix} \right\} \Big/ N \quad (9)$$

The table 2 shows the six (6) first ones resultants patterns of that classification that presented larger occurrence.

Table 2: Patterns classification for the proposed method

Pattern	*Weight*	*Occurr*	*%*
P1={1,0,0,0,0,0,0,1,0,0,0,1}	0.76448	22	6.27
P2={1,0,0,1,0,0,0,0,0,0,0,0}	0.75427	10	2.85
P3={1,0,0,1,0,0,0,0,0,0,0,1}	0.75309	10	2.85
P4={1,0,0,1,0,0,0,1,0,0,0,1}	0.75095	14	4.00
P5={1,0,1,0,0,0,0,1,0,0,0,1}	0.73243	11	3.14
P6={0,0,0,0,0,0,1,0,0,0,0,0}	0.68329	16	4.57

Notice that the patterns classification doesn't determine the most characteristic pattern of the disease. That classification only informs the most frequent patterns considering the presence or absence of a symptom. That strategy allows diagnosing symptoms patterns that match or not with the disease.

Obtaining Discriminant Rules

Starting from the classification of Table 2, discriminant rules of the expressed type were built in (10). The rules allow determining how much a pattern characterizes patients sick and not sick.
Considering $Pk = \{pk_1, pk_2, ..., pk_r\}$ where $r = 1...symptoms$ as being the group of patterns found in the database, the discriminant rules are given by:

$$\begin{aligned} &\forall(S^i), where\ s^i_j = pk_j \forall\ j = 1,..,N \Rightarrow \\ &reqPac(S^i)[w1] \wedge freqSick(S^i)[w2] \end{aligned} \quad (10)$$

The rule should be read as: *for any patient "i" that possesses the symptoms pattern P has the probability of occurrence of w1 and the probability of being sick of w2.*
Applying the patterns of the Table 2 to the symptoms database and considering the expression (10) was obtained the following discriminant rules:

$$\begin{aligned} &\forall(S^i), where\ S^i = P1 \Rightarrow \\ &freqPac(S^i)[6.27\%] \wedge freqSick(S^i)[40{,}91\%] \end{aligned}$$

$$\begin{aligned} &\forall(S^i), where\ S^i = P4 \Rightarrow \\ &freqPac(S^i)[4.00\%] \wedge freqSick(S^i)[64.29\%] \end{aligned}$$

Obtaining the patients' profiles

In way to identify patients' groups that characterize the disease, were applied the discriminant rules in the patients' database. The Table 3 shows data of patients for each rule.

4 RESULTS

In the Table 1, was observed that 63% of the patients presented the symptom s_1 and 70% of them had nephrolithiasis. On the other hand, 37% of the patients didn't present the symptom s_1, and 96% of them are sick. In the same way, 42% of the patients presented the symptom s_5, and 80% of them are also sick patients. On the other hand 58% of the patients didn't present the symptom s_5, and 80% of them were sick patients.
The previous results do not allow us extract conclusions if the symptom characterizes or not the disease. For that reason was necessary a classification of patterns considering the presence and the absence of symptoms. The Table 2, shows the most important patterns, where can be observed that the symptoms s_1, s_3, s_4, s_7, s_8 e s_{12} have a certain relevance. The symptom s_1 corresponds to HiperCalcidio, s_3 corresponds to HiperUricidio, s_4 corresponds to HipoCitratidio, s_8 corresponds to HFLRFamily (genetic inheritance) and s_{12} corresponds to Hematuria.
Through the Table 2, is observed that the pattern *P4* almost have all the characteristic symptoms for the disease. Its representation was expressed through the rule:

$$\begin{aligned} &\forall(S^i), where\ S^i = P4 \Rightarrow \\ &freqPac(S^i)[4.00\%] \wedge freqSick(S^i)[64.29\%] \end{aligned}$$

Table 3. Patients' characteristics for each pattern

Pattern	Sex (%)		Race (%)					Current age		Start-Disease	Weight (kg)	Height (cm)
	M	F	1	2	3	4	5	$\bar{x}$	$S(x)$			
1	62	38	44	6	37	0	13	18.5	1.2	0 – 9	26.60	130
2	63	37	55	9	36	0	0	13.5	5.8	0 – 10	24.81	126
3	45	55	82	4	10	4	0	13.1	4.0	2 – 10	22.22	122
4	40	60	30	10	50	0	10	15.4	6.4	0 – 6	29.00	112
5	80	20	50	0	40	10	0	14.4	4.9	0 – 10	19.30	114
6	46	54	55	15	15	0	15	11.6	3.4	2 – 10	23.84	116

which is possible to conclude: from the 351 analyzed patients, 4% (14 patients) presented the pattern *P4* and 64.29% (9 patients) of them were sick. For the pattern *P1*, 6.27% (22 patients) of the patients had this pattern, and 40.91% (9 patients) were sick.

$$\forall(S^i),\ where\ \ S^i = P1 \Rightarrow$$

$$freqPac(S^i)[6.27\%]\wedge freqSick(S^i)[40{,}91\%]$$

Notice in Table 3, that the fields <Current Age>, <Start-Disease>, <Weight> and <Height> are not important in relationship to <sex> and <Predominant Race>. For the P1 rule is possible to say that patient of masculine sex and of white race is able to acquire the disease. This last conclusion is very superficial due to data sets that referring to the feeding they are not available. For future works will be considered nutritious habits and areas where the patients live.

5 CONCLUSIONS

In this work, an application of the KDD methodology for knowledge discovery in the database for nephrolithiasis was applied. The problem of the medical diagnosis for this type of disease, especially in children, is a difficult work of diagnose and advancing. For that reason the database of the sick and not sick was considered. That influence in a positive way in the medical sector and it affects positively in the medical diagnosis.

The work was divided in three stages: the first stage corresponds to the identification of patterns in clinical and laboratorial symptoms, to characterize the disease in sick and not sick patients. For the second stage discrimination rules were used on the patterns of discovered symptoms. The last stage corresponds to a discrimination process, of the patient's profile, based on the obtained rules.

The limitation of the number of data sets is a great obstacle for the success of Data Mining algorithms and for the KDD methodology. It is possible to discover patterns (knowledge) valid for a group specific of data sets. The current researches should give attention to that subject. The database used in this work contains 351 records of patients with clinical and laboratorial symptoms. In spite of the "little" amount of data, is probably the only database in the net of hospitals in the Belo Horizonte city in Brazil.

It is necessary methodologies that allow extract knowledge in database with few records. In this paper are proposed some heuristic techniques capable to work with few data. If the KDD methodology and Data Mining try to discover gold, perhaps be in the hour of we begin to seek that gold under our feet, in smaller and more accessible database than the great databases. Another great challenge can be in front of us.

REFERENCES

Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. 1996. Advances in Knowledge discovery and data mining. Menlo Park, CA: AAAI Press/MIT Press.

Pyle, D. 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc. San Francisco, California.

Bojarczuk, C; Lopes, H. Freitas, A. Data Mining with Constrained Syntax Genetic Programming: Applications in Medical Data Set. Intelligent Data Analysis in Medicine and Pharmacology. IDAMAP01, London, UK, September, 4th, 2001.

Werner, J.; Fogarty, T. Genetic Programming Applied to Severe Diseases Diagnosis. Intelligent Data Analysis in Medicine and Pharmacology. IDAMAP01, London, UK, September, 4th, 2001.

Jones, M.D. 14 Powerful Techniques for Problem Solving. Times Books Random House Inc. 1998