Pharmaceutical Data Search by Decision Trees

Djamila Benhaddouche
Faculty of Mathematics and Computer Science
University Of Science and Technology of Oran Mohamed Boudiaf Usto-MB
Oran, Algeria
Email: Benhaddouche.djamila [AT] gmail.com

Abstract— **The number of information concerning the drugs that any professional of health must control in practice and the transformations which they undergo, make the regulation or the administration of drugs difficult for a pregnant woman. Techniques of excavation of data were developed to lead a model of classification of data according to precise criteria. One of the most used of is the technique of the decision trees, a method making it possible to predict the membership of an individual to a class according to his characteristics; it is based primarily on the relevant attributes of the data base of the field to which it is applied. In our case classification of managed drugs or not with the pregnant woman will be done according to quarters of the pregnancy. The results of this technique will help the professionals of health to take a decision, to make a good regulation, to decrease the accidents related to the catch of inadequate drugs at the period of pregnancy with less risks for the child.**

Keywords-component; data mining; supervised learning; decision trees; segmentation,; fragmentation.

I. INTRODUCTION

The term Data Mining is often employed in order to designate the whole of the tools which permit tothe user to accede to the company's data, to analyse them.[1] Here we will restrict the term of Data Mining to tools which aim the generation of rich information form company's data notably historical data, to discover implicit modals in data.

They can permit for example to a shop to extract profiles of costumers and also typical purchasing to forecast future sales. It permits to increase the value of data that are containing in Data Warehouse.2

II. TECHNIQUE OF THE DECISION TREES

The decision's trees are used within the framework of discovery of directed knowledge. [5]They are very powerful tools mainly used for classification, description or the estimate. The principle of operation is as follows: to explain a variable, the system seeks the criterion the most determinant and cutting the population

in under populations having the same entity of this criterion. Each under population is then analyzed like theinitial population. The returned model is easy to understand and the founded rules are very explicit. This system is appreciated very much. The decision trees are a structure which is often used to represent knowledge. [5]

III. THE C4.5 ALGORITHM [6]

Algorithm developed by J.Ross Quinlan into 93 [3]. The interior version was called ID3, and is still usedin some products. It is always supposed that the language of representation consists a certain number ofattributes. These attributes can be binary, qualitative or continuous. For the continuous attributes, we use theheuristic ones which make it possible to discretize them. We use for this; the statistical criteria which make itpossible to achieve the two following goals: a number of classes not too significant and a good distribution enter the various classes. We can for example use the function entropy to achieve these goals.

IV. EXPERIMENTAL STEPS

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. Data preparation

The preparation is an important stage, if it is not of primary importance process of retrieval of knowledge starting from data. It is a question as well as possible of defining the individuals and there presentation used for the training. The quality of the model of classification depends largely on the quality of this preparation.

This preparation could not be done without the assistance of the experts of this field; it consists in obtaining the adequate data for the realization of our objective. These data are structured in fields. From our data base we identified the following drug. [7].

1. Drugs managed with the pregnant woman in first quarter of its pregnancy.

2. Drugs managed with the pregnant woman in the second quarter of its pregnancy.

3. Drugs managed with the pregnant woman in the third quarter of its pregnancy.

4. Drugs managed with the pregnant woman until the second quarter of its pregnancy.

5. Drugs managed with the pregnant woman as from the second quarter of its pregnancy.

6. Drugs managed with the pregnant woman through her pregnancy.

7. Drugs which cannot be managed with the pregnant woman in all the period of her pregnancy

B. Enrichment

After we have studied the information contained in our data base, we noted that some of the necessary Information for the achievement of our objective did not exist in this data base. Let us note that one can Solve problems only if one has the necessary data. For this reason we have to make an enrichment of the data Base by adding the attributes which are corresponding to our needs.

We initially made a library research for the collection and the comprehension of the data, which were presented to our experts (doctors and pharmacists), they approved the choice of the data and took part with Their preparation. The attributes with which we have to enrich our data base are as follows: adverse effects, Counters indications, toxicity, risk, use.

C. Dateggregation

According to our study which was undertaken close to our experts, so the drugs that can be managed with the pregnant woman are only those whose route of administration is one of the following ways: oral, rectal, sublingual, and inject able Since our data base contains all possible shapes of drugs (tablets, galls,syrup, injectable solution…) with the result that the Form attribute takes a great number of discrete values,
we made a regrouping of these values in order to obtain an number of reasonable values.[8] We gathered them according to their routes of administration and according to whether they are managed with the woman in girded or not, which means that we have renown the Form field in Route Of administration whose possible passage are: oral, rectal, sublingual, and injectable for the drugs concerning the pregnant woman, for the other routes of administration we have to attribute to these drugs other values.[7][8].

D. Cleaning

Our data come from a warehouse, which means that they were already cleaned, but because of the enrichment we carried out on this basis and knowing the fact, it is extremely possible that there are doubled blooms, errors of seizes, or even missing information we made a cleaning for management of what follows: The doubled blooms can appear awkward because they will give more importance to the repeated values.
The errors of seizures represent disturbed data, they are awkward in the case or they would hide a repetition.
When a field does not contain any information, one is in the case of a missing data. Our data base contains much missing data but that is not a problem for us because the algorithm that we have choosed (C4.5) [3]has the advantage of managing the missing data by replacing them by the majority value of the field in question in the data base.

E. Data mining

The size of our initial population, which is useful for the training, is of 165 recordings. The algorithm is launched. The results of our study are as follows:

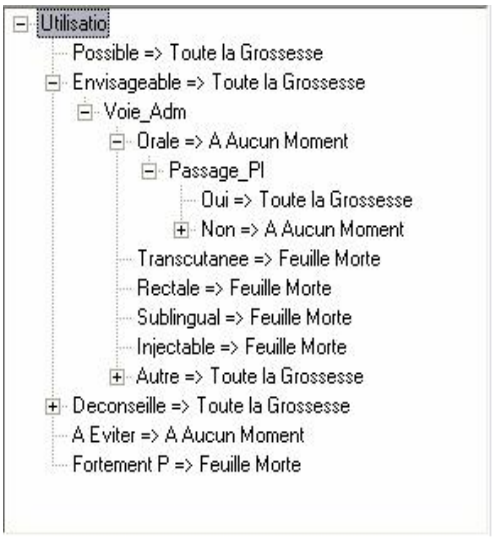

Figure 1. The following figures represent the decision tree obtained for our population of 165 recordings.

Figure 2. . this figure is a progressive development on the higher level of the nodes of the tree.

Figure 3. . this figure is a progressive development on the higher level of the nodes of the tree

After the stage of the construction of the tree, the stage comes from the pruning, which consists in eliminating under trees with an aim of minimizing the real error of the tree obtained (post-pruning).

After the construction, the tree is transformed into rules of deduction that is able to be interpreted with more facility.

These rules are form If prémisse1 and prémisse2 and...Prémisse N then result.

Results following for the first branch: If use = possible then all the pregnancy.If utilization=envisage able and sees administration=oral then No Moment.

Moreover, it is the same for the remainder's branches of the decision tree.

Unless Our data come from a warehouse, which means that they were already cleaned, but because of the enrichment we carried out on this basis and knowing the fact, it is extremely possible that there are doubled blooms, errors of seizes, or even missing information we made a cleaning for management of what follows: The doubled blooms can appear awkward because they will give more importance to the repeated values.

The errors of seizures represent disturbed data, they are awkward in the case or they would hide a repetition.

When a field does not contain any information, one is in the case of a missing data. Our data base contains much missing data but that is not a problem for us because the algorithm that we have choosed (C4.5) [3] has the advantage of managing the missing data by replacing them by the majority value of the field in question in the data base.

V. CONCLUSION

The medication of database that we used contains very interesting information but badly exploited; the addition of new data as we carried out only for our problems increased the interest of this database as well as the need of tools of excavation of data for the extraction of new useful and relevant information for the decision-making. [2]

The use of such tool of excavation of data made more possible to solve the problem of taking decision that concern the regulation of drugs for the pregnant woman according to the period of her pregnancy by decreasing the risks of accident of an inadequate regulation.[7]

We think we will improve our results by introducing the CART method. [4].

Our essential objective for this study and to help the expert is to say the obstetrician-gynecologists and not to decide for them. Most of the regional pharmacovigilance centers are a real guide to prescribing and assessing drug risk during pregnancy. It is intended for the daily practice of gynecologists-obstetricians, midwives and general practitioners, but also for their initial or continuing training. [8].

REFERENCES

[1] René Lefébure and Gilles Venturi - Eyrolles, 1998 "Le Data Mining"

[2] Administrative guide to the list of medications- February 2020

[3] J.R. Quinlan, C4.5 programs for machine learning Morgan Kaufmann Publishers, livres Google

[4] Comparison of C5.0 & CART Classification algorithms using pruning technique International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 4, June - 2012, ISSN: 2278-0181.

[5] J.R. QUINLAN, Induction of Decision Trees, 1986, Machine Learning 1:81-106.

[6] Johan Baltié, DataMining : ID3 et C4.5, Promotion 2002, S.C.I.A. specialization School for computer science and advanced techniques.

[7] Drug Prescriptions in Pregnant Women: A Descriptive Analysis in the Department of the Aube Hérique A , Proy M-O , Truchi L , Verroust P. 2001.

[8] Medication and Pregnancy: Prescribing and Assessing the Risk. Edited by Annie-Pierre ,Jonville-Béra, Thierry Vial. Book 2012.