Mining Healthcare Data: The Case of an Endoscopic Thoracic Sympathectomy Dataset

MARIBEL YASMINA SANTOS, DIANA GONÇALVES
Algoritmi Research Centre
University of Minho, Guimarães, PORTUGAL
maribel@dsi.uminho.pt, pg13247@uminho.pt

JORGE CRUZ
Medicine Faculty, University of Lisbon, Lisboa, PORTUGAL
costacruzjorge@gmail.com

Abstract: - The process of knowledge discovery in databases aims at the discovery of associations within data in a dataset. Data Mining is a central step of this process corresponding to the application of algorithms for identifying patterns in data. This paper presents the particular case of analysis of a dataset containing data associated with 227 patients submitted to an endoscopic thoracic sympathectomy, a treatment for primary palmar hyperhidrosis. Primary hyperhidrosis is characterized by an excessive sweating that appears as a consequence of a disorder of the sympathetic autonomous nervous system. The results achieved show an overall improvement of the patients’ quality of life, mainly associated with their emotional state.

Key- Words: - Knowledge discovery in databases, data mining, decision trees, primary hyperhidrosis, endoscopic thoracic sympathectomy

1 Introduction
Primary hyperhidrosis is a disorder of the sympathetic autonomous nervous system that affects around 1% of the global population [1]. It is characterized by an excessive sweating of the face, palms, armpits and feet, either in all of these locals or in some of them. This excessive sweating provokes several problems to the individual, who sees his/her life completely influenced by this disorder.

Endoscopic upper-thoracic sympathectomy has been considered the treatment of choice for primary palmar hyperhidrosis [2]. This surgery is a minimal invasive procedure of thoracic sympathetic blockage and consists of the bilateral ablation of the second and third thoracic sympathetic ganglions, affecting the sympathetic nervous outflow to the arms and elsewhere [3-4].

Being a definitive treatment for this disorder, the results of the surgery have revealed a high degree of patient satisfaction [2]. To show the increase of the quality of life and the incidence of complications and side effects after the surgery, this paper shows the analysis of data collected from 227 patients. Although several studies have been conducted in order to evaluate the improvement in the patients’ health condition [1-6], the analysis presented in this paper exploits a new perspective in data analysis, using data mining algorithms to analyze the available data.

After the analysis of the collected data, it was verified an improvement of the overall health condition of the patients, with a significant change in their emotional state. The incidence of compensatory hyperhidrosis was analyzed as it is one of the major complains of the patients after the surgery. These results are presented later on in this paper.

This paper is organized as follows. Section 2 presents an overall overview of the knowledge discovery process and the main steps associated with it. Section 3 gives an overall overview of the collected data and the questionnaire used in the data collection process. Section 4 presents the results achieved using the Clementine Data Mining System® for data analysis. Section 5 concludes with some remarks about the work undertaken.

2 Knowledge Discovery in Databases
Knowledge Discovery in Databases (KDD) is a process that aims at the discovery of associations within datasets. Data Mining is a central step of this process. It corresponds to the application of algorithms for identifying patterns from data without the additional steps of the knowledge discovery
process such as incorporation of appropriate domain knowledge and interpretation of results [7].

The steps of the KDD process include data selection, data treatment, data pre-processing, data mining and interpretation of the results. This process is interactive, because it requires the user participation, and iterative, because it allows for going back to a previous phase and then proceeding with the knowledge discovery process. The five steps [8] of this process are now briefly described:

Data Selection. This step allows the selection of the relevant data needed for the execution of a defined data mining task.

Data Treatment. This phase is concerned with the cleaning up of the selected data, allowing the treatment of errors and the definition of strategies for dealing with missing data fields.

Data Pre-Processing. This step makes possible the reduction of the sample available for analysis. Two tasks can be carried out here: i) the reduction of the number of rows or, ii) the reduction of the number of columns. For those, data can be generalized using the defined hierarchies or attributes with continuous values can be transformed into discreet values according to the defined classes.

Data Mining. Several algorithms can be used for the execution of a given data mining task. In this step, several algorithms are evaluated in order to identify the most appropriate for the defined task. The selected one is applied to the relevant data in order to find patterns in data.

Interpretation of Results. The interpretation of the discovered patterns aims at evaluating their utility and importance with respect to the application domain. In this step it can be realized that relevant attributes were ignored in the analysis, thus suggesting that the process should be repeated.

Different tasks can be performed in the knowledge discovery process and several techniques can be applied for the execution of a specific task. Among the available tasks are classification, clustering, association, and summarisation. KDD applications integrate a variety of data mining algorithms. The performance of each technique (algorithm) depends upon the task to be carried out, the quality of the available data and the objective of the discovery. The most popular data mining algorithms include neural networks, decision trees, association rules and genetic algorithms [9].

In this paper, decision trees are used to identify and test classification models, which can be used to understand the evolution in the quality of life of patients submitted to an endoscopic thoracic sympathectomy.

3 Available Data

The study presented in this paper collected and analyzed data from 227 patients. An adapted version of the SF-6D questionnaire was made available through the Internet to 504 patients that were submitted to an upper-thoracic sympathectomy in two Portuguese health institutions: the Hospital Santa Maria in Lisbon and the Instituto Cardiovascular de Lisboa, also in Lisbon. A letter containing a description with the objectives of this work, a set of instructions, a login and a password (for an on-line answering) was sent to each patient. From this population of 504 patients, an answering rate of 48% was obtained (30 questionnaires did not reach their destination due to wrong addresses or changes of addresses).

The SF-6D questionnaire is based on the short form health survey questionnaire (SF-36), a promising instrument for measuring health perception in a general population [10]. The SF-36 instrument integrates eight health concepts: physical functioning, bodily pain, role limitations due to physical health problems, role limitations due to personal or emotional problems, emotional well-being, social functioning, energy/fatigue, and general health perceptions. The SF-36 questionnaire was already used in several studies for evaluating the quality of life in individuals submitted to an endoscopic thoracic sympathectomy for primary hyperhidrosis treatment [5-6].

A validated Portuguese version of the SF-6D was used [11]. This version was upgraded to include specific information about the disorder in study and the complications and side effects verified as consequence of the surgery. Each patient answered to two questionnaires. One associated to his/her health condition prior the surgery and the other related with his/her health condition in a post-surgery scenario. Personal information included in the questionnaires, like name, address, birth date and gender of the patients was of optional answering. Information related to the surgeon’s name, surgery date and patient’s age at surgery was also of optional answering. Although optional, the majority of the patients provided this information allowing the characterization of the population in study.

In terms of the patients’ characteristics, the female patients are predominant, with 159 individuals. The dataset includes 66 male patients. Two patients did not mention their gender. With respect to the age at surgery, the sample data set includes individuals with a wide range of ages, from 10 to 52 years old. However, it is on the range of 20 to 39 years old that most of the surgeries were done.
All the contacted patients had available an online platform (through the Web) to answer the questionnaires (the prior and the post-surgery). The collected data were verified in order to see if any errors were present. As answers were limited to a fixed set of valid options, no errors were found in data. A few cases of missing data fields were detected, mainly on personal information. In these cases, all fields were marked with the “Unknown” or “Not know” labels. For some missing data fields on continuous attributes, the average value was calculated and assigned to the missing values [9]. After the data verification process, the collected data were loaded into a data mart in order to support the analysis process. The data mart model was defined with the decision process in mind: i) know the characteristics of the patients, ii) verify the occurred side effects and complications, and iii) see the evolution of the patients’ quality of life.

4 Data Analysis

The data collected in this study was stored in a data mart that integrates six main vectors of analysis, each one of them represented by one of the fact tables of the model [12]. The constellation schema integrates the following fact tables: Hyperhidrosis, Hyperhidrosis Drawbacks, Compensatory Hyperhidrosis, Emotional State, Side Effects, and Quality of Life.

The Hyperhidrosis fact table allows the storage of all the information associated with the hyperhidrosis incidence verified in the patient before and after the surgery. The Hyperhidrosis Drawbacks is used to characterize the beginning of the disorder in the patient and when he/she started to feel uncomfortable with it. The Compensatory Hyperhidrosis table stores the incidence of compensatory hyperhidrosis in the patients both one month and six months after the surgery. The Emotional State table characterizes the emotional feelings of the patients, namely how they are affected by this disease. The Side Effects table stores the temporary and/or permanent side effects verified in the patients as consequence of the surgery. The Quality of Life table receives the answers to the questions that are related with the SF-6D questionnaire and the quality of life measure calculated from the given answers. The quality of life measures were obtained using the SF-6D utility made available by Professor John Brazier [10].

The data analysis process starts by analysing the improvement in the emotional state of the patients. The available data integrate the several types of emotional states (Incomprehension feeling concerning the disease, Fear of compliments, Insecurity, Need to hide the disease, Socialization difficulties, Shame of talk about the disease and Isolation), the grade of incidence of each emotional state (Inexistent, Irrelevant, Expressive, Excessive or Not know), some personal information associated with the patients like Gender and Age at Surgery, and the information about the patient’s situation, namely, if he/she is in a prior surgery or post-surgery scenario (attribute Questionnaire). Table 1 shows an extract of the available data for the emotional state.

Table 1 - Characterization of the emotional state

As the available data was stored in a data mart, data was cleaned and treated before the loading process. Due to this fact, the knowledge discovery process starts by splitting the available data into two datasets, the training dataset and the test dataset, and by using the C5.0 algorithm for the induction of a decision tree. Fig. 1 presents the stream implemented in the Clementine Data Mining System® for the execution of these tasks. Besides the splitting of the available data, nodes (1-in-3) and 1-in-3, with a partition of 67% for training and 33% for testing, the stream also integrates a web node that analyses the Emotional State Grade and the Questionnaire attributes. After the identification of the decision tree for this specific task, the obtained model, node Questionnaire with a diamond shape, is analyzed with an Analysis node to verify the confidence of the model.

Fig. 1 - Stream for the analysis of the emotional state
The use of the web node allowed the identification of the graph pictured in Fig. 2. A web graph shows the relationships that exist among the values of the different attributes and also the strength of each relationship. As thinner is the link between the values, as weak is the relationship. Strong relations are represented by bold lines. Analyzing the graph in Fig. 2 it is possible to see that before the surgery the most representative emotional state is Expressive, and that after the surgery it is the Inexistent grade. This graph started to show the improvement of the emotional state of the patients.

Fig. 2 - Web node for the analysis of the emotional state grade

In the data mining phase, and as already mentioned, the C5.0 algorithm was used for the induction of a decision tree. The obtained tree is shown in Fig. 3. It includes four rules for the prediction of the scenario in which the patient is, attending to his/her Emotional State Grade and Emotional State Type. If the Emotional State Grade is one of Excessive, Expressive or Not know, the patient is in a prior-surgery scenario in 90.57% of the cases. If the Emotional State Grade is Inexistent then the patient is in a post-surgery scenario in 82.80% of the cases. If the grade is Irrelevant, it is necessary the analysis of the Emotional State Type to get a conclusion. The types Fear of compliments and Incomprehension feeling concerning the disease lead to a nearly tie scenario between the prior and the post-surgery cases. In the cases of Insecurity, Isolation, Need to hide the disease, Shame of talk about the disease and Socialization difficulties, the patient is in a prior-surgery scenario in 63.24% of the cases.

After the analysis of the obtained decision tree, its performance was verified in the test dataset. Fig. 4 shows an accuracy of 81.48% when classifying unknown cases.

Fig. 3 - Decision tree for the emotional state
Fig. 4 - Analysis of the decision tree for the emotional state

Continuing with the data mining analyses, the incidence of compensatory hyperhidrosis was verified as this is the major complain of the patients after the surgery. In this analysis, the Compensatory Hyperhidrosis Type, the Compensatory Hyperhidrosis Grade, the Gender, the Age at Surgery and the Questionnaire attributes were analyzed. The type can be Abdominal, Legs or Dorsal, while the grade is one of Unsupportable, Excessive, Moderate, Irrelevant, Inexistent or Not know. Concerning the age, in this dataset the age of the patients ranged from 10 to 52 years old. Table 2 shows an extract of the available data.

Table 2 - Data for the analysis of the compensatory hyperhidrosis

Fig. 5 presents the stream implemented for the analysis of the incidence of compensatory hyperhidrosis. Once again, the available data was divided into the training and testing datasets with a percentage of 67% and 33%, respectively.
The generated web node allowed the verification of the incidence of the several grades by the different compensatory hyperhidrosis types. Analyzing Fig. 6 we verify that the majority of the patients feel a Moderate compensatory hyperhidrosis in the three analyzed body parts: Abdominal, Legs and Dorsal. Although its low incidence, the Unsupportable grade is verified in some of the patients.

The target attribute of the decision tree for the verification of the incidence of the compensatory hyperhidrosis was the patients’ Gender. This option was motivated by the fact that male (M) and female (F) present different complains concerning this surgery side effect. The obtained decision tree (Fig. 7) shows that the grades Inexistent, Irrelevant, Moderate and Not know are verified by female in 74.65% of the cases. For the Excessive and Unsupportable grades, the Age at Surgery is required to classify the gender of the patients. Following the tree it is possible to see that younger males tend to be more affected by these grades, in opposition to female, which present a different incidence in what age concerns.

The confidence of the obtained model is presented in Fig. 8, with an accuracy of 72.77%. As this is not a high value, it is recommended the collection of more data and its consequent analysis to see the impact on the obtained results.
5 Conclusion

This paper presented the analysis of a dataset related with 227 patients submitted to an endoscopic thoracic sympathectomy for the treatment of primary hyperhidrosis. Being a definitive treatment for this disorder of the sympathetic autonomous system, the results of the surgery have revealed a high degree of patient satisfaction. Through the use of Data Mining algorithms it was possible to verify the improvement of the emotional state of the individuals, the incidence of compensatory hyperhidrosis (one of the main side effects of the surgery) and the improvement of the overall quality of life of the patients. Future work includes the collection of more data to get a deeper understanding of the effects of this surgery and to obtain models with higher accuracies in their prediction capabilities.

References:


