Informatics in Medicine Unlocked 12 (2018) 120–127

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

The prediction of good physicians for prospective diagnosis using data mining

Nfongourain Mougnutou Rémy[a,*], Tekinzang Tedondjio Martial[a], Tayou Djamegni Clémentin[b]

[a] Department of Mathematics and Computer Science, FS, University of Ngaoundere, Ngaoundere, Cameroon
[b] Department of Mathematics and Computer Science, University of Dschang, Dschang, Cameroon

ARTICLE INFO

Keywords:
Data mining
Open data
Logistic regression
Multidisciplinary diagnosis

ABSTRACT

This work provides a predictive model for selecting the most appropriate health care practitioners, particularly physicians, to diagnose a patient. In the context of a multidisciplinary diagnosis, this paper provides a data mining model to identify a specialist physician who can participate in such a diagnosis and thus reduce the risk of errors. First, the model identifies the specialists who can diagnose a patient. Second, the model uses the calculated probabilities to provide a ranking of specialist physicians capable of making a good diagnosis. This ranking can be used to construct a group of specialists who can participate in the multidisciplinary diagnosis. A sample of 58177 patients (52% women) consulted by 11 different medical specialists was extracted from the SPARCS database. The work is based on the analysis of open health data, specifically, diseases that keep patients stable. The result of the data mining is a multinomial logistic regression model. The 10-fold cross-validation results indicate that the model provides good predictive capability for the selected data, with an average accuracy, sensitivity, specificity, and precision of 80%, 79%, 97.3%, and 82.8%. Our results show that a patient's characteristics influence the selection of a physician. In conclusion, we assert that all selected specialists are able to diagnose the patient and that some specialists have a greater ability to diagnose the disease than do others.

1. Introduction

Medicine has undergone changes with the rapid progress of science and new approaches by physicians, which have resulted in modern medicine. Despite these advances, diagnostic errors persist in medicine [27]. To reduce the risk of incorrect diagnosis by a single physician, a *multidisciplinary approach to diagnosis* can be considered [7].

In the case of multidisciplinary diagnosis, multiple actors from different fields collaborate to provide a single diagnosis. The group work required by multidisciplinary diagnosis makes it possible not only to obtain value but to facilitate the identification and analysis of the causes of the patient's problem. However, an expert's opinion may be considered to be relevant only in his or her areas of expertise. In other words, one physician's opinion about a patient's problem may be more welcome than the opinion of another physician. This raises the *problem of choosing the physicians* to participate in multidisciplinary diagnosis, as illustrated in Fig. 1.

The objective of this work is to construct a model that can *describe the relationship between the perspective of a good diagnosis provided by a physician and a patient's profile*. To achieve this objective, an analysis of health data was conducted, and a statistical representation of the physician-patient relationship was developed. This paper presents a model for selecting medical experts based on several characteristics of a patient. The selected specialists can thus participate in a multidisciplinary diagnosis. To the best of our knowledge, this is the first study of its kind on physician prediction based on a patient's profile using data mining.

2. Background

2.1. Data mining

Data mining is the art of finding information or knowledge in a large amount of data. Like statistics, data mining is becoming increasingly common in companies and organizations that want to extract relevant information from their databases, which they can use for their own needs [31]. Data mining tasks can, in general, be classified as tasks of *description* and *prediction* [30], [24]. To understand the discovery goal, it is vital to understand the difference between descriptive and predictive tasks.

Data mining technology is applied in an increasing variety of fields. In descriptive data mining, the goal is to produce a descriptive

* Corresponding author.
E-mail addresses: nfongourain@yahoo.fr (N.M. Rémy), tekinzang@yahoo.fr (T.T. Martial), dtayou@yahoo.com (T.D. Clémentin).

https://doi.org/10.1016/j.imu.2018.07.005
Received 2 March 2018; Received in revised form 13 July 2018; Accepted 29 July 2018
Available online 02 August 2018

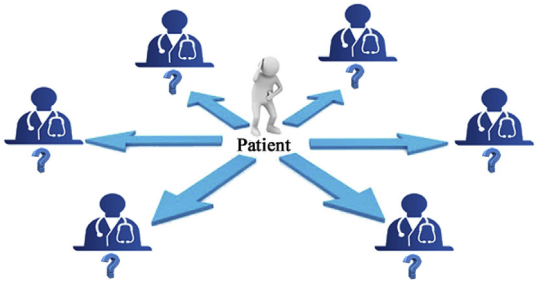

Fig. 1. The problem of selecting physicians.

approximation or model of the process that generates the data. The goal of prediction is to find a model to estimate the values of future cases.

Medical data mining is an essential component of clinical decision support systems. Data mining can extract hidden information in the data of the medical domain and exploit it as patterns for clinical diagnosis [14], [33]. Similar to [14] and [33], most of the work in the literature focuses on patients' diseases and disregards physicians. The aim of this work is to propose a model with the probability of a patient's profile that suits the modality of a physician, in other words, to predict the physicians who can provide a good diagnosis based on the patient's characteristics.

2.2. Predictive data mining

In most data mining applications, a target variables on which we will learn is necessary [25]. A predictive model can be understood as data learning. In this context, we also need to know the value of the target variables for a set of examples (i.e., patient records).

2.3. Algorithms

Access to electronic patient records (EHRs) opens new possibilities for medical data mining. Many different supervised machine learning algorithms can be used for analysing datasets. Some of the techniques of data mining that are successfully used in healthcare today are *decision trees (DTs), artificial neural networks* and *logistic regression*.

DTs are one of the most powerful and popular tools to extract information. They also have several advantages [4]. A considerable asset of a DT is that it has the advantage of being a highly interpretable model that represents a set of rules. However, other machine learning algorithms, such as the support vector machine (SVM) [19], may provide better accuracy yet builds less interpretable models.

Artificial neural networks [16] are derived from the analysis and information processing of the human brain. They represent knowledge as a network of units, or neurons, that are present in the brain. ANNs have been successfully used in applications in clinical medicine, such as diagnosis in medical images [8]. The method has been tested on several problems and compared with several existing methods, and it obtained performance comparable to that of SVM. However, compared to SVM, artificial neural networks have much longer execution times and do not explain their results.

Logistic regression (LR) is one of the most widely used methods for statistical modelling of binary response variables. LR predicts the probability of the target variable, denoted by p. The target variable can have a value of 1 (success) or a value of 0 (failure, $1 - p$). LR has been used extensively in the medical and social sciences [20]. Multinomial logistic regression (MLR) is an alternative to binomial logistic regression [15]. Multinomial logistic regression has an advantage; it does not assume a linear relationship among the dependent variable and each independent variable. MLR is used in situations in which there is no ordering of K values of dependent variables, which is the case in our study. In this work, the dependent variable is nominal and consists of more than 02 categories. We focus on multinomial regression to estimate the probability of selecting each of the categories and the effect of independent variables on the outcome.

2.4. Open data

The world of data is becoming increasingly competitive every day, as observed in terms of volume, variety and value. Open data adds richness and new dimension to data warehouses and analysis to unlock new forms of innovation [3]. The sharing and opening up of data make it possible to make essential data available online and to improve the analysis of many decision-makers, thus improving the ability to make more informed decisions in various sectors including medicine [26]. This therefore means the creation of large sets of reference data shared by all stakeholders and the encouragement of the development of several high value-added services. Open data means that these data are available for access, exploitation and reuse by any interested party (companies, scientists, etc.).

This work was performed using medical information from the Health Statewide Planning and Research Cooperative System (SPARCS) database. SPARCS is a database of patient characteristics, diagnoses, treatments and services of patients whose lesional and/or functional status is considered to be stable (e.g., angina). The French Clinical Classification of Emergency Patients (CCMU) [32] commonly used for care prioritization from level 1 to level 5 is:

- *Level 1*: Clinical condition considered to be stable. Simple clinical examination. Abstention of complementary diagnostic or therapeutic procedures.
- *Level 2*: Level 1 and decision of additional diagnostic procedure (e.g., blood test).
- *Level 3*: Clinical condition may worsen without any life-threatening prognosis.
- *Level 4*: Life-threatening risk without starting immediate resuscitation procedures.
- *Level 5*: Vital prognosis engaged involving starting resuscitation procedures.

In this work, all selected diseases have one thing in common, they maintain the patient's clinical condition and/or functional prognosis stable.

3. Design of the predictive model

3.1. Data understanding and data preparation

3.1.1. Patient problem

The clarification of the patient's problem involves performing the patient's history. The collection of a patient's demographic data is the starting point of a patient's history. The next step is the development of the patient's profile and the patient's chief complaint [29]. Demographic data and the patient's profile are important because they provide a representation of the patient and his or her condition, including, age and gender. This information can also help to identify other medical problems. A patient's profile is a summary of the patient's characteristics and problems that lead to the patient's current condition.

In this paper, we want to identify suitable physicians who can work together to solve a patient's problem. The process of selecting physicians is illustrated in Fig. 2. The core of the process is the predictive model that must be able to predict which physicians can provide a good diagnosis based on a patient's profile.

3.1.2. Preprocessing

The SPARCS 2014 database used in this work has more than 1,000,000 observations and 39 variables in its raw state, which includes many missing, redundant and irrelevant values. All variables not correlated with the objectives of the study were removed. No grouping was