Applied Soft Computing 47 (2016) 424–437

Contents lists available at ScienceDirect

Applied Soft Computing

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

Solving the problem of incomplete data in medical diagnosis via interval modeling

Andrzej Wójtowicz, Patryk Żywica, Anna Stachowiak, Krzysztof Dyczkowski*

Department of Imprecise Information Processing Methods, Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form 12 April 2016
Accepted 18 May 2016
Available online 24 May 2016

Keywords:
Missing data
Uncertainty
Aggregation
Medical diagnosis
Decision-making

ABSTRACT

This paper presents an approach to making accurate and high-quality decisions under incomplete information. Our comprehensive approach includes interval modeling of incomplete data, uncertaintification of classical models and aggregation of incomplete results. We conducted a thorough evaluation of our approach using medical data for ovarian tumor diagnosis, where the problem of missing data is commonly encountered. The results confirmed that methods based on interval modeling and aggregation make it possible to reduce the negative impact of lack of data and lead to meaningful and accurate decisions. A diagnostic model developed in this way proved better than classical diagnostic models for ovarian tumor. Additionally, a framework in R that implements our method was created and is available for reproduction of our results. The proposed approach has been incorporated into a real-life diagnosis support system – OvaExpert.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The aspect of data uncertainty is studied intensively in many contexts and scientific disciplines, including medicine. Many different forms of uncertainty in data have been recognized: one comes from conflicting or incomplete information, as well as from multiple interpretation of some phenomenon; another arises from lack of well-defined distinctions or from imprecise boundaries. Functioning under uncertainty and ignorance is an everyday experience of many practitioners, and is impossible to eliminate completely. For example, in medical practice it has been shown [1,2] that the collection of complete data by a physician during examinations can be highly problematic due to the technical limitations of the healthcare institution, the high costs of a medical examination or the high risk of deterioration in a patient's health after a potential examination. The lack of data hinders the use of traditional models for diagnosis support, and there is therefore an urgent need to solve this problem.

One of the possible approaches to managing incompleteness of data is to exploit well-established methods from the field of data imputation (see Ref. [3]). Undoubtedly in many research areas such an approach is sufficient. However in medical applications, where human life is at stake, it is not so clear whether we can introduce new data which may be subject to small but significant error. Another option is to develop a new model specially dedicated to incomplete data. However, the multiplicity of already existing models makes it difficult to select the right one among them, and consequently physicians are confused and refrain from using any of them. Adding yet another diagnostic model would increase complexity in modeling and computation even further. For these reasons we explore an entirely different path. The main idea is to construct a general method that makes it possible to adapt and integrate existing and well-established diagnostic methods to make them usable with incomplete data.

A direct motivation of our work was the need to support gynecologists in diagnosis of ovarian tumor, including in the case of incomplete data. This type of cancer is particularly difficult to diagnose, and its mortality rates have remained high for many years [4]. The main problem is to determine whether a tumor is malignant or benign based on two groups of parameters: data from medical history (e.g. age, weight, number of pregnancies) and diagnostic data (e.g. blood markers, ultrasonography). The research problem therefore boils down to a binary classification problem.

There are several well-known models in ovarian tumor diagnostics. Some of them are created by individual research units, such as the Alcazar model and SM; others by organizations (incorporating a number of research centers), such as IOTA LR1. The majority are scoring models and models based on logistic regression. These models attain different levels of effectiveness [5,6], generally high on internal data but very often much lower during external

* Corresponding author. Tel.: +48 61 8295402; fax: +48 61 8295315.
E-mail address: chris@amu.edu.pl (K. Dyczkowski).

http://dx.doi.org/10.1016/j.asoc.2016.05.029
1568-4946/© 2016 Elsevier B.V. All rights reserved.

validation. Different models use different patient attributes, and collecting all of them may be costly and problematic. Moreover, these models are not prepared for the case where some of the data in the patient description are missing. Recently, IOTA developed the first model that is able to handle missing value of one attribute [7]. In this paper we want to propose general method for handling missing values. The importance of the completeness and quality of medical data was recently highlighted in [8].

As a result, the ability to diagnose – called *diagnosability* or *decisiveness* – of these models may be low in many practical situations. So far, all research in this field were made on complete data sets. In consequence the problem of data incompleteness is not well investigated although it is currently discussed in medical community [9]. Furthermore, unlike in other classification problems, there is no clearly defined and widely accepted indicator of the quality of such a diagnostic model. The most commonly used metrics are the area under the ROC curve (AUC), accuracy, and sensitivity. However, these do not reflect all of the aspects of the problem analyzed here; in particular they do not take into account the level of diagnosability.

In our approach we were able to turn some of the abovementioned drawbacks into assets, so as to achieve a solution to the problem of incomplete data. During the research we noticed that, since the models use different attributes, they complement one another, allowing better decisions to be made. However, gynecologists were not yet able to take advantage of this fact. To change this, we developed the idea of creating a decision support system that would integrate knowledge derived from a number of models, and provide it in an accessible way to the doctor.

We have developed OvaExpert, a specialist diagnostic system to support gynecologists, including those less experienced, in the proper differentiation of tumors. The results presented in this paper answer problems encountered during work on the OvaExpert system. The system is currently being intensively tested at a number of medical centers. The main objective of the system is to make accurate decisions despite a lack of data. This is achieved by interval modeling of incomplete information. The use of the diversity of diagnostic models allows us to increase the efficiency of diagnosis by aggregating knowledge from many sources. To this end, we implemented a number of aggregation operators and conducted a set of tests to verify how those operators act on real-life data, both complete and incomplete. By sharing our work through GitHub, we enable other researchers to verify our results and to reuse our code for their own purposes. We believe that our results may prove valuable not only in ovarian tumor diagnosis, but also in other classification tasks in which the problem of missing and incomplete data is faced.

The remainder of the paper is organized as follows. In Section 2 we present details of our approach to dealing with data incompleteness, including uncertaintification of patient descriptions and diagnostic models, and methods of information aggregation. Section 3 describes an evaluated dataset as well as the evaluation procedure. In Section 4 the results of our experiments are presented and discussed. Section 5 emphasizes the significance of our results by giving a short introduction to their application in the OvaExpert system. Conclusions appear in Section 6.

2. Proposed approach

The main objective of our approach is to enable effective decision-making, in spite of missing data. The most obvious approach, based on imputation, is not feasible here for many reasons. First of all we were limited by the very small number of cases that could be used as a prior knowledge for effective imputation. Besides, as has already been mentioned in the previous section, imputing the results of diagnostic tests, even though it may be correct from a statistical point of view, can lead to significant diagnostic error. Imputation can serve as a convenient way of carrying out statistical analyses of a dataset or a classifier, but it must be clearly stated that imputed data are not the real one so it may be hazardous to use them for making a diagnosis for one particular patient. This issue was widely discussed in a recent book by Hatch [9]. Our primary objective was not to make an illusion of operating on complete data. We want a doctor to be aware of the incompleteness of the knowledge about a patient's state and rather to suggest no diagnosis then the wrong one. Finally, our ultimate goal is to develop a general method that deals not only with totally incomplete (missing) data but also with data complete only to some extent (interval data), for which imputation is not the answer.

In our research, we adopted the following two assumptions. Firstly, we accept a state in which a diagnostic model does not return any diagnosis. This should not happen too often, but in the most difficult diagnostic cases (or if a significant part of attributes is missing) it may be the only option. Secondly, we do not intend to create new diagnostic models.

Instead, we enable the use of existing models under missing and incomplete data. We base our research on available regression and scoring models. Theoretical example of such model as well as our approach is illustrated in the following subsections (Examples 1–4). More details about the models are given in Section 3.1.

2.1. Interval modeling

In a classical approach, a patient is modeled as a vector $\mathbf{p}$ in a space P. Let $D_1, \ldots, D_n$ be real closed intervals denoting domains of attributes that describe patients. We define a set P in the following way $P := D_1 \times \ldots \times D_n$. Then, a vector $\mathbf{p}$ that describes a patient has the form $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, where $p_i \in D_i$.

A diagnostic model can be formalized as a function $m : P \to [0, 1]$. The values returned by the function indicate confidence as regard the malignancy of a tumor, and are interpreted in the following way:

- $m(\mathbf{p}) \geq 0.5$ – diagnosis toward malignant (higher values represent higher confidence);
- $m(\mathbf{p}) < 0.5$ – diagnosis toward benign (lower values represent higher confidence).

Observe that the situation where $m(\mathbf{p}) = 0.5$ is resolved toward malignancy.

Example 1. For the sake of simplicity, in this example we assume that the patient is described only by two attributes, namely patient's age and one cancer antigen test. We define the domains of these attributes as $D_1 = [0, 100]$ and $D_2 = [0, 1500]$. Consider the following two patients: $\mathbf{p}^A = (35, 100)$ and $\mathbf{p}^B = (60, 1200)$. Let $m_1 : P \to [0, 1]$ be a simple example diagnostic model defined by

$$m_1(\mathbf{p}) = 0.0025p_1 + 0.0005p_2.$$

Now we can easily see that according to diagnostic model m_1 patient A should be diagnosed as benign ($m_1(\mathbf{p}^A) = 0.138$) and patient B as malignant ($m_1(\mathbf{p}^A) = 0.75$).

The existing diagnostic models operate on complete patient data. In order to represent missing values we have to add a special element (in practice commonly denoted by NA) to the domain of each attribute. Thus patient is now described by a vector $\mathbf{p} = (p_1, \ldots, p_n)$, where $p_i \in D_i \cup \{\text{NA}\}$. A major disadvantage of this approach is the need to introduce a new, separate value to represent missing values. This value cannot be handled natively by the original diagnostic models, which in turn leads to an inability to make any

Download English Version:

https://daneshyari.com/en/article/494633

Download Persian Version:

https://daneshyari.com/article/494633

Daneshyari.com