Engineering Applications of Artificial Intelligence 76 (2018) 67–79

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Liver fibrosis diagnosis support using the Dempster–Shafer theory extended for fuzzy focal elements

Sebastian Porebski [a,*], Piotr Porwik [b], Ewa Straszecka [a], Tomasz Orczyk [b]

[a] Institute of Electronics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[b] Computer Systems Department, Institute of Computer Science, University of Silesia, Bedzinska 39, 41-200 Sosnowiec, Poland

ARTICLE INFO

Keywords:
Liver fibrosis
Diagnostic rule extraction
Dempster–Shafer theory
Medical diagnosis support

ABSTRACT

Classifiers are used in a variety of applications, among them the classification of medical data. Their efficiency depends on the quality of training data, which is a disadvantage in the case of medical data that are often imperfect (e.g. incomplete, imbalanced, uncertain). Moreover, numerous classifiers are black-boxes from the perspective of diagnosticians who perform the final diagnoses. These drawbacks degrade the potential usefulness of classifiers in diagnosis support. A rule-based reasoning may overcome these mentioned limitations. We introduce both a rule selection and a diagnosis support method based on the Dempster–Shafer and fuzzy set theories. The theories can manage an interpretation of incomplete and imbalanced data, imprecision of medical information and knowledge uncertainty. The usefulness of the method will be proven on a test case of liver fibrosis diagnosis. The liver fibrosis stage is difficult to recognize even for experienced physicians. The diagnosis of the liver state by an invasive biopsy is ambiguous and dependent on its finite precision. Therefore, knowledge-based methods are being sought to reduce the need of invasive testing. We use a real medical database related to patients affected by hepatitis C virus to extract knowledge. The database has missing and outlying values and patients' diagnoses are uncertain. The proposed methods provide simple diagnostic rules that are helpful in this study of liver fibrosis and in processing deficient data. The greatest benefit and novelty of the approach is the ability to assess three stages of fibrosis in a non-invasive way, whereas other medical tests allow to detect only the last stage, i.e. the cirrhosis.

1. Introduction

Nowadays, we observe significant development of classification methods used to cope with the constantly growing amount of information (Porwik et al., 2016; Orczyk and Porwik, 2015; Foster et al., 2014; Sheikhpour et al., 2017; Karakatič and Podgorelec, 2016). Many are also applied in medical diagnosis support (Esfandiari et al., 2014). When a medical data set is available, classifiers are able to learn to distinguish healthy and ill patients and after that, efficiently predict diagnosis for a new consulted patient. However, finding the best individual classifier for a diagnostic problem is difficult, both from a statistical and a computational point of view (Woźniak et al., 2014). Moreover, there are two main issues that cast doubt on the usefulness of classifiers in the medical domain.

Classification methods are highly dependent on the quality of the training data set (Burduk, 2014; Orczyk and Porwik, 2015; Porwik et al., 2016; Straszecka, 2010). Ideal training sets should be complete and appropriate for the patients' population. These conditions are difficult to fulfill especially for medical data, since they are usually imperfect (Esfandiari et al., 2014). Medical information quality can be degraded when measurements are incorrectly performed or inadequate (Khaleghi et al., 2013). Data credibility can be also questioned when the data come from various sources or refer to the past. Medical data are often incomplete e.g. because of changes in diagnostic procedures or the inability to perform an examination. In medical records incomplete data often co-exist with imbalanced data sets. The lack of data values affect the whole data set while an inequality of data number occurs for some diagnosis sets. For example, only the most necessary tests are performed for a patient because of costs and onerousness, and the number of patients with hepatic fibrosis culminating in cirrhosis is much smaller compared to the number of patients with other liver fibrosis stages. Mentioned data deficiencies can spoil the training step of a classifier and consequently its performance in final diagnosis determination. Even if we put off the problem of database quality, classifiers rarely focus on interpretability of their performance. Thus, diagnosticians often treat them as black-box

* Corresponding author.
E-mail address: sebastian.porebski@polsl.pl (S. Porebski).

https://doi.org/10.1016/j.engappai.2018.09.004
Received 18 January 2018; Received in revised form 31 July 2018; Accepted 8 September 2018

solutions and they are skeptical about their conclusions (Berner, 2010). It is clear that medical diagnosis support using classifiers must deal with the mentioned difficulties.

A diagnosis rule extraction is an idea to model medical knowledge that is "hidden" in the available medial data records. Knowledge should be represented by interpretable rules. Conditional rules, known as if-then rules, can be a basis of cooperation between an engineer and a diagnostician. The rules, verified by an expert or left unchanged after the extraction can be used to evaluate a diagnosis for an unknown case. In this paper, we extract diagnostic rules that are used in diagnosis support based on the Dempster–Shafer theory (Dempster, 1968; Shafer, 1976) boosted with the fuzzy set theory (Zadeh, 1999). The Dempster–Shafer theory allows combining information from different sources (Leung et al., 2013; Hui et al., 2017) and this method is resistant to incomplete data sets (Straszecka, 2006). It has been already used in the medical domain (Dezert et al., 2012; Paul et al., 2012; Gui et al., 2017) as well as linked to set theories (Yen, 1990; Khatibi and Montazer, 2010; Xiao et al., 2012; Wang et al., 2016). In the framework of this theory diagnostic rules can be represented by focal elements. Thus, the focal elements are pieces of medical knowledge crucial for a particular diagnosis problem (Straszecka, 2010). It is quite common to use fuzzy sets in the focal element to model the imprecision of medical data (Nguyen et al., 2015; Pourpanah and Lim, 2016; Romero-Córdoba et al., 2017). However, the proposed approach is substantially different. Its main advantage is considering focal elements as individually defined fuzzy sets instead of exploiting an equivalence between fuzzy sets and the basic probability assignment (Yen, 1990; Xiao et al., 2012; Li et al., 2015; Wang et al., 2016). In our method imprecision of symptoms and uncertainty of diagnostic rules are distinctly modeled by fuzzy sets and basic probability, which gives our approach an advantage over standard classifiers.

In Xiao (2018) a belief in a diagnostic hypothesis is evaluated using the Dempster–Shafer theory and fuzzy soft sets. This interesting method allows for changing the set of symptoms that are used in diagnosis, but requires using the preference matrix. Dependence on this matrix is inconvenient for expert analysis and it does not provide clear diagnostic rules. Given that knowledge extraction, along with effective diagnosis, is the principle goal of the present approach, fuzzy soft sets are not helpful. It must be also noted that the classical combining of basic probability assignments that is used in Xiao (2018) in the Dempster–Shafer theory that is used in the mentioned method, sometimes bring results far from expert's estimation (Straszecka, 2016).

Rule extraction must be particularly accurate to avoid obtaining conflicting knowledge when deficient data (either incomplete and imbalanced) are considered. In this paper we present rule selection algorithms that allow finding the diagnostic rule set that provides best relation between symptoms and diagnosis (Porebski and Straszecka, 2018). In our approach, we evaluate symptoms importance individually for each diagnosis by means of the Matthews Correlation Coefficient (MCC) (Powers, 2011). This is the main difference in the evaluation of symptoms between our approach and other typical feature selection methods that usually do not estimate symptoms in such a detailed manner.

Advantages of the proposed approach will be proven for the real problem of liver fibrosis diagnosis. We present exploratory knowledge extraction results from real medical liver fibrosis data. These are data of patients that are affected by HCV (hepatitis C virus) and participate in therapy. Blood tests and a biopsy examination with METAVIR (meta-analysis of histological data in viral hepatitis) scale are done for the patients (Hytiroglou et al., 1995). The biopsy result is treated as the reference diagnosis hence the diagnosis is as certain as this examination. Unfortunately, even an experienced diagnostician sometimes has trouble in distinguishing the intermediate states of liver diagnosis in METAVIR scale (Porwik et al., 2016; Krawczyk et al., 2013). Therefore, the reference diagnosis is uncertain. A significant part of the data set is missing and outlying values can be also observed. Hence, the knowledge extraction needs an approach that can deal with these types of data deficiency. As a result we present a simple and readable rule set that can be used in non-invasive diagnosis. It can also form the basis of a diagnosis support system. In the discussion section, comparative studies are presented and discussed. These results explore the performance of other approaches tested on the considered database as well as different non-invasive methods of liver fibrosis diagnosis.

Summarizing, the main contributions of the paper are the following:

- we introduce an algorithm for problems where input data are noisy — both imbalanced and incomplete, deficiencies that can be observed in various stored medical databases,
- the Dempster–Shafer theory is applied in such a way that it allows separate analysis of the symptom imprecision as well as the medical knowledge uncertainty. Its application improves the classification accuracy and provides clear diagnostic indications,
- the new rule extraction algorithm has been experimentally validated on an anonymous, real medical database. Experiments prove that it outperforms state-of-the-art methods of medical diagnosis support.

2. Methods

2.1. Model of the medical data

Let us denote by l the diagnosis index and by C the number of all considered diagnoses, so $l = 1, \ldots, C$. Then, we can describe the medical data set related to the lth diagnosis in the following way:

$$\boldsymbol{X}^{(l)} = \begin{bmatrix} x_{11}^{(l)} & x_{12}^{(l)} & \cdots & x_{1j}^{(l)} & \cdots & x_{1r}^{(l)} \\ x_{21}^{(l)} & x_{22}^{(l)} & \cdots & x_{2j}^{(l)} & \cdots & x_{2r}^{(l)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1}^{(l)} & x_{i2}^{(l)} & \cdots & x_{ij}^{(l)} & \cdots & x_{ir}^{(l)} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n_l1}^{(l)} & x_{n_l2}^{(l)} & \cdots & x_{n_lj}^{(l)} & \cdots & x_{n_lr}^{(l)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{p}_1^{(l)} \\ \boldsymbol{p}_2^{(l)} \\ \vdots \\ \boldsymbol{p}_i^{(l)} \\ \vdots \\ \boldsymbol{p}_{n_l}^{(l)} \end{bmatrix} \tag{1}$$

where $x_{ij}^{(l)}$ is the value of jth symptom of the ith patient. The number of all symptoms in the data set is denoted by r. The number of all patients related to the lth diagnosis is denoted by n_l. The data in the matrix $\boldsymbol{X}^{(l)}$ (1) can be represented by row vectors. In this case medical data of the ith patient ($\boldsymbol{p}_i^{(l)}$) can be expressed as follows:

$$\boldsymbol{p}_i^{(l)} = [x_{i1}^{(l)}, x_{i2}^{(l)}, \ldots, x_{ij}^{(l)}, \ldots, x_{ir}^{(l)}], \tag{2}$$

whereas the jth symptom of all n_l patients assigned to the lth diagnosis is the column vector of the matrix $\boldsymbol{X}^{(l)}$ and will be denoted as:

$$\boldsymbol{x}_j^{(l)} = [x_{1j}^{(l)}, x_{2j}^{(l)}, \ldots, x_{ij}^{(l)}, \ldots, x_{n_lj}^{(l)}]^\top. \tag{3}$$

2.2. Detection of outlying values

Outlying values are a characteristic phenomenon of medical data. A patient with unusual symptoms is not a rare case. Hence, we are not able to judge whether the outlying value is the result of a mistake or it is a real abnormality. Regardless its origin, in our research outlying values are detected and ignored to ensure an effective rule set extraction. Usually the Z-score is applied to detect outliers (Iglewicz and Hoaglin, 1993). The outlying value is recognized among the jth symptom values (3) for one diagnosis, i.e. outlying $x_{ij}^{(l)}$ is found in $\boldsymbol{x}_j^l$. The Z-score is calculated in the following way (Iglewicz and Hoaglin, 1993):

$$z(x_{ij}^{(l)}) = \frac{0.6745(x_{ij}^{(l)} - \tilde{x}_j^{(l)})}{MAD(\boldsymbol{x}_j^{(l)})}, \tag{4}$$