Expert Systems With Applications 91 (2018) 36–45

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A multi-level classification framework for multi-site medical data: Application to the ADHD-200 collection

Sarah Itani[a,b], Fabian Lecron[c,*], Philippe Fortemps[b]

[a] Fund for Scientific Research - FNRS (F.R.S.-FNRS), Brussels, Belgium
[b] Department of Mathematics and Operations Research, Faculty of Engineering, University of Mons, Mons, Belgium
[c] Department of Engineering Innovation Management, Faculty of Engineering, University of Mons, Mons, Belgium

ARTICLE INFO

Article history:
Received 2 May 2017
Revised 19 July 2017
Accepted 26 August 2017
Available online 30 August 2017

Keywords:
Clinical decision support systems
Decision trees
Multi-level approach
Attention Deficit Hyperactivity Disorder (ADHD)

ABSTRACT

Recently, the culture of sharing medical data has emerged impressively, reducing significantly the barrier to the development of medical research accordingly. As open-access large datasets result from this significant initiative, data mining techniques can be considered for the development of interpretable expert systems to help in diagnosis. However, the collaborative effort of information gathering yields heterogeneous databases because of technical and geographical factors. Indeed, on the one hand, the harmonization of protocols for data collection is still missing. On the other hand, cultural and social factors impact locally both the epidemiology and etiology of a given disease. Ignoring these factors could weaken the credibility of studies based on multi-site data. Thereby, our work tackles the development of computer-aided diagnosis systems relying on heterogeneous data. For such a purpose, we propose a multi-level approach (inspired by multi-level statistical modeling) based on decision trees (in the sense of machine learning). This framework is applied on the public ADHD-200 collection for the study of Attention Deficit Hyperactivity Disorder (ADHD).

1. Introduction

During the last decade, the sharing of large-scale medical data appears to be a growing trend encouraged by the scientific community. Several medical databases were launched publicly to address different health concerns (Church, 2005; Di Martino et al., 2014; Esfandiari, Babavalian, Moghadam, & Tabar, 2014; Hunter et al., 2005; Ihle et al., 2012; Kerr, Lau, Owens, & Trefler, 2012; Milham, Fair, Mennes, Mostofsky et al., 2012; Mueller et al., 2005). Some initiatives aim not only at sharing databases, but also software tools to manage information at best (Milham, 2012). Such a culture of data sharing is valuable to the research sphere (Mennes, Biswal, Castellanos, & Milham, 2013; Piwowar, Becich, Bilofsky, Crowley et al., 2008; Ross, 2016; Ross & Krumholz, 2013). Indeed, access to data is henceforth facilitated, especially to researchers not having the medical information on hand usually. Among others, specialists of computer science and mathematics can contribute their expertise at the technological level notably. The interaction of medicine with other disciplines is made possible therefore. Moreover, with the sharing of large-scale data, a same issue is approached in different ways worldwide, so research is enriched and accelerated. At last, a common framework of open-access data encourages local and international research centers to make their own databases available online. This virtuous circle multiplies the mass of available information, and the quality of studies is improved accordingly.

In particular, opening such access to data allows to focus on the explanation of some diseases/troubles through the detection of physiological foundations and/or typical symptoms. The advent of data mining allows to meet these needs, and to develop expert systems for the purpose of aid in diagnosis (Esfandiari et al., 2014; Parvathi & Rautaray, 2014). Such a decision support system should preferably be *interpretable*, i.e. able to show how a diagnosis is acquired (Lavrač, 1999), and *readable*, through assessment criteria making sense (Wagholikar, Sundararajan, & Deshpande, 2012)

The efforts deployed to gather data collaboratively are undeniably outstanding, nevertheless, they yield databases whose use is challenging. Indeed, the harmonization of data collection protocols is still missing (Milham et al., 2012). Across sites, data acquisition differs in terms of equipment calibration, experimental conditions, and sampling methodologies (Abraham et al., 2017). There exist also variations in the strategies that are used to process medical images (Abraham et al., 2017). Besides, the influence of cultural and social factors on both the epidemiology and etiology of dis-

* Corresponding author.
E-mail addresses: sarah.itani@umons.ac.be (S. Itani), fabian.lecron@umons.ac.be (F. Lecron), philippe.fortemps@umons.ac.be (P. Fortemps).

http://dx.doi.org/10.1016/j.eswa.2017.08.044
0957-4174/

eases is established since the past century (Landy, 1977; Link & Phelan, 1995; Trostle, 2005). These factors of disparities are added to the natural heterogeneity of medical data, caused by the different natures of information available on a patient (interviews, phenotype, scans) and their interpretation by physicians (Cios & Moore, 2002; Wasan, Bhatnagar, & Kaur, 2006). Hence, multi-site medical data remain in reality a patchwork of subsets that cannot be merged into a single dataset without any adaptation. As inconsistency may arise from the use of heterogeneous medical data, studies generally focus on subsets withdrawn from these large databases (Abraham et al., 2017). The collective attempts to share large open-access data are therefore partially rewarded: admittedly, the access to medical data is easier, but the open-access databases remain not fully exploited.

In our study, we attempt to tackle the development of interpretable and readable diagnosis support models able to cope with a medical multi-site database. We propose an application to the ADHD-200 collection (Milham et al., 2012), an example of stable and recent release of multi-site medical database launched for the study of Attention Deficit Hyperactivity Disorder (ADHD). The main contributions of our work are exposed below.

- After the study of Abraham et al. (2017), we propose another way to deal with the issue of heterogeneous multi-site medical databases. The work of Abraham et al. (2017) proposed two cross-validation strategies that enable to acquire models less sensitive to the heterogeneity of multi-site medical databases. In our work, instead of questioning the validation phase to address this issue, we propose a rethinking of the learning process to develop models able to help in diagnosing a disease/trouble, under a novel *hierarchical* spirit.
- In previous works, to predict a diagnosis, hierarchical systems were set up to practice classification in two steps: (1) to dissociate healthy and pathological cases and (2) to detect the type of pathology. The methodologies lied on the development of either two classifiers per dataset (site) or two classifiers across the whole database. The novelty of our proposal is related to the hybrid nature of our hierarchical classifier where *intra-site* and *inter-site variabilities* play both a role at different levels to deliver a final robust diagnosis. To the best of our knowledge, such a procedure is innovative and allows to take into consideration geographical parameters in contrast to the work of Abraham et al. (2017). We will name our proposition as a *multi-level approach* since it is inspired by the theory of multi-level analysis.
- The work presents new results as concerns the ADHD-200 collection. In particular, we do not consider domain-specific features to solve the classification task, but rather features having demonstrated success in other domains. The results show that these *meaningful features* provide interesting interpretations for helping in diagnosing ADHD.

In Section 2, we will expose our case study, before developing the materials and methods of this work in Section 3. The results of our study will be presented and discussed in Section 4. Finally, we will conclude this paper in Section 5.

2. A case study: attention deficit hyperactivity disorder

Approximately five to seven percent of children and teenagers are likely to be confronted one day with Attention Deficit Hyperactivity Disorder (ADHD)[1]. Also affecting adults, this mental trouble is characterized by inattention, and/or hyperactive-impulsive attitudes. Generally, people with ADHD have to deal with a reduced self-control impairing notably their ability to express serenely their feelings, to the detriment of their social and professional daily life.

Today, the subject's environment (mainly parents and teachers) constitutes almost the only source of information that practitioners have to make a diagnosis, unmistakably subjective. Research is still ongoing to better understand the physiological bases of the trouble.

In 2011, researchers from various fields of expertise were challenged to propose an objective assessment of ADHD in the context of the ADHD-200 contest (Milham et al., 2012). As a working basis, a multi-site medical database, called the ADHD-200 collection, was released online[2]. The database includes clinical (phenotypic) and neuroimaging data (resting-state functional and structural magnetic resonance images) on altogether 947 patients. Typically Developing (TD) and ADHD affected patients are included in the database. ADHD cases are expressed in three types: Inattentive (ADHD-I), Hyperactive-Impulsive (ADHD-HI) and a Combination of both types (ADHD-C). Eight sites contributed to the collection of data: Peking University (PU), Kennedy Krieger Institute (KKI), NeuroImage (NI), New-York University (NYU), Oregon Health & Sciences University (OHSU), University of Pittsburgh (Pitt.U), Washington University in St. Louis (WU) and Brown University[3]. Tables 1 and 2 present the subjects' distribution according to multiple criteria as regards both training and test sets (see also Milham et al., 2012). Let us note that the ADHD-HI type (Hyperactive-Impulsive) was not predicted as the associated population is very low in the training set.

The ADHD-200 collection can be qualified as a heterogeneous medical dataset as it includes instances of various geographical origins whereas social factors influence the local prevalence of ADHD (Russell, Ford, Rosenberg, & Kelly, 2014); the subsets were not collected according to a common protocol (Bellec et al., 2017); the gender representativeness as well as the healthy and pathological cases proportions can significantly vary according to sites.

Table 1
Summary of the ADHD-200 training data.

Site	Age	Gender		TD	ADHD	Total
		F	M			
PU	8–17	52	142	116	78	194
KKI	8–13	37	46	61	22	83
NI	11–22	17	31	23	25	48
NYU	7–18	77	145	99	123	222
OHSU	7–12	36	43	42	37	79
Pitt.U	10–20	43	46	89	0	89
WU	7-22	28	33	61	0	61
Total		290	486	491	285	776

Table 2
Summary of the ADHD-200 test set.

Site	Age	Gender		TD	ADHD	Total
		F	M			
PU	8-15	19	32	27	24	51
KKI	8-12	1	10	8	3	11
NI	13–26	13	12	14	11	25
NYU	7–17	13	28	12	29	41
OHSU	7–12	17	17	28	6	34
Pitt.U	14–17	2	7	5	4	9
WU	–	–	–	–	–	–
Total		65	106	94	77	171

[1] http://www.adhd-institute.com/burden-of-adhd/epidemiology/.

[2] The dataset is available at http://www.fcon_1000.projects.nitrc.org/indi/adhd200/.

[3] Brown University data were discarded in this work as the results of diagnosis are not supplied for this site.