COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 122 (2015) 1–15

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Dealing with inter-expert variability in retinopathy of prematurity: A machine learning approach

V. Bolón-Canedo[a,*], E. Ataer-Cansizoglu[b], D. Erdogmus[b], J. Kalpathy-Cramer[c], O. Fontenla-Romero[a], A. Alonso-Betanzos[a], M.F. Chiang[d]

[a] Department of Computer Science, Universidade da Coruña, A Coruña, Spain
[b] Cognitive Systems Laboratory, Northeastern University, Boston, MA, USA
[c] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Charlestown, MA, USA
[d] Departments of Ophthalmology & Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

ARTICLE INFO

Article history:
Received 26 February 2015
Received in revised form 6 May 2015
Accepted 7 June 2015

Keywords:
Inter-expert variability
Clinical decision-making
Feature selection
Machine learning
Classification
Retinopathy of prematurity

ABSTRACT

Background and objective: Understanding the causes of disagreement among experts in clinical decision making has been a challenge for decades. In particular, a high amount of variability exists in diagnosis of retinopathy of prematurity (ROP), which is a disease affecting low birth weight infants and a major cause of childhood blindness. A possible cause of variability, that has been mostly neglected in the literature, is related to discrepancies in the sets of important features considered by different experts. In this paper we propose a methodology which makes use of machine learning techniques to understand the underlying causes of inter-expert variability.

Methods: The experiments are carried out on a dataset consisting of 34 retinal images, each with diagnoses provided by 22 independent experts. Feature selection techniques are applied to discover the most important features considered by a given expert. Those features selected by each expert are then compared to the features selected by other experts by applying similarity measures. Finally, an automated diagnosis system is built in order to check if this approach can be helpful in solving the problem of understanding high inter-rater variability.

Results: The experimental results reveal that some features are mostly selected by the feature selection methods regardless the considered expert. Moreover, for pairs of experts with high percentage agreement among them, the feature selection algorithms also select similar features. By using the relevant selected features, the classification performance of the automatic system was improved or maintained.

Conclusions: The proposed methodology provides a handy framework to identify important features for experts and check whether the selected features reflect the pairwise agreements/disagreements. These findings may lead to improved diagnostic accuracy and standardization among clinicians, and pave the way for the application of this methodology to other problems which present inter-expert variability.

* *Corresponding author*. Tel.: +34 981 167000.
E-mail address: vbolon@udc.es (V. Bolón-Canedo).

http://dx.doi.org/10.1016/j.cmpb.2015.06.004

1. Introduction

Retinopathy of prematurity (ROP) is a disease affecting low-birth weight infants, in which blood vessels in the retina of the eye develop abnormally and cause potential blindness. ROP is diagnosed from dilated retinal examination by an ophthalmologist, and may be successfully treated by laser photocoagulation if detected appropriately [1]. Despite these advances, ROP continues to be a major cause of childhood blindness in the United States and throughout the world [2]. This is becoming increasingly significant in middle-income countries in Latin America, Eastern Europe and Asia because these countries are expanding neonatal care, yet have limited expertise in ROP. In addition, the number of infants at risk for ROP throughout the world is increasing dramatically because of improved survival rates for premature infants [3], while the availability of adequately-trained ophthalmologists to perform ROP screening and treatment is decreasing [4].

An international classification system was developed during the 1980s, and revised in 2005, to standardize clinical ROP diagnosis [5]. One key parameter of this classification system is called "plus disease", and is characterized by tortuosity of the arteries and dilation of the veins in the posterior retina. Plus disease is a boolean parameter (present or absent), and is the most critical parameter for identifying severe ROP. Numerous clinical studies have shown that infants with ROP who have plus disease require treatment to prevent blindness, whereas those without plus disease may be monitored without treatment. Therefore, it is essential to diagnose plus disease accurately and consistently.

However, high levels of inconsistency among experts when diagnosing ROP have been demonstrated [6,7]. Inter-expert variability in clinical decision making is an important problem which has been widely studied in the literature for several decades [8]. Much of this previous work has examined inter-expert variability in the interpretation of ophthalmic images [9,6,10,11]. There are also studies which focus on the variability in diagnosis of acute diseases such as prostate cancer [12], breast cancer [13], melanoma [14], papillary carcinoma [15], and polycystic ovary disease [16]. Although there is a broad range of studies on analysis of inter-expert variability, few of them focus on investigating its underlying causes [17–20].

Understanding the causes of disagreement among experts is a challenging problem. In the cognitive process during clinical diagnosis, some features may be considered more important by certain experts than by others. If two experts consider different sets of features during diagnosis, then we might expect to see a strong disagreement between them. Hence, such a feature-observer analysis enables us to understand the underlying causes of inter-expert variability.

In this work, we propose a methodology for investigating the important features for the experts when diagnosing ROP, with the final aim of building automated diagnosis systems. The proposed system makes use of *feature selection*, which is a machine learning technique employed to detect the most important features for a given classification task [21]. After selecting the useful features for each expert, we carry out a similarity analysis to see if the selected features can reflect the disagreement among experts. Finally, we propose an approach to build automated diagnosis tools applying machine learning techniques. The contributions of this paper are, (i) use and comparison of various feature selection algorithms to understand the underlying causes of inter-expert disagreement, (ii) a similarity analysis to validate whether feature selection results are consistent with the disagreement among experts, and (iii) the construction of an automatic diagnosis system that makes use of the feature selection results and similarity analysis findings.

In our previous work [20], we proposed a method to investigate whether there are groups of observers who decide consistently with each other and if there exist important features these experts mainly focus on. The previous approach involved a hierarchical clustering of the experts using a pairwise similarity based on mutual information between the diagnostic decisions. Next, we performed an analysis to see the dependence between experts' decisions and image-based features which enabled us to qualitatively assess whether there are popular features for the group of observers obtained through clustering. Different than our previous study, in this work (i) we provide an in-depth analysis to find important features for each expert using various feature selection algorithms, (ii) we validate the feature selection results performing a quantitative similarity analysis between the selected features and the experts' agreement (i.e. we expect to select the same features for expert pairs with a high degree of agreement), and (iii) we build an automated classification system considering the analysis results and compare different classification algorithms.

The remainder of this paper is organized as follows: Section 2 explains the research methodology, and Section 3 details the problematics of ROP diagnosis. Finally, Section 4 reports the experimental results, and Section 5 describes the discussion of the main findings and conclusions.

2. Research methodology

In order to develop automatic systems that can support clinicians in the diagnosis of ROP, it is necessary to extract the knowledge from the medical experts. However, as discussed before, there is a high degree of disagreement among experts, and the reasons behind this disagreement are not clear. This paper proposes a methodology to understand the causes of inter-expert variability in ROP diagnosis, as a step toward extracting the necessary knowledge to build an automatic diagnosis tool.

A four-step methodology is thus applied, as illustrated in Fig. 1. First, the problem needs to be analyzed to check if disagreement among experts exists. Second, several feature selection methods are applied to discover which features are the most important to each individual expert. Third, a similarity analysis is performed to check if, for experts with a high ratio of agreement, the feature selection methods also select similar features. Finally, the classification performance

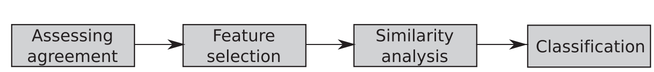

Fig. 1 – Steps of the research methodology.