Expert Systems With Applications 113 (2018) 147–160

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data

Tommy Hielscher[a,*], Uli Niemann[a], Bernhard Preim[b], Henry Völzke[c], Till Ittermann[c], Myra Spiliopoulou[a]

[a] Otto-von-Guericke University Magdeburg, Department of Technical and Business Information Systems, Universitätsplatz 2, Magdeburg D-39106, Germany
[b] Otto-von-Guericke University Magdeburg, Department of Simulation and Graphics, Universitätsplatz 2, Magdeburg D-39106, Germany
[c] University Medicine Greifswald, Institute for Community Medicine, Walter Rathenau Str. 48, Greifswald D-17475, Germany

ARTICLE INFO

Article history:
Received 19 September 2017
Revised 11 June 2018
Accepted 2 July 2018
Available online 2 July 2018

Keywords:
Subpopulation discovery framework
Constraint-based subspace clustering
Cohort study data
Hepatic steatosis
Goiter

ABSTRACT

Objective: We propose an intelligent system that assists epidemiology experts in analysing the data of a population-based epidemiological study, in identifying relevant variables for an outcome and subpopulations with increased disease prevalence, and in validating the findings concerning variables and subpopulations in a further, expert-specified cohort. At present, the study of an outcome on a population-based cohort is hypothesis-driven, i.e. the expert must specify the variables to be studied. Our approach rather operates in a data-driven, semi-automated way, enabling the expert to identify variables of relevance and generate hypotheses on them.

Methods: Our system DIVA supports the *D*iscovery, *I*nspection and *VA*lidation of subpopulations with increased prevalence of an outcome, without requiring parameter tuning. DIVA takes as input the cohort of an epidemiological population-based study with *all* variables specified in the study's protocol, as well as inputs from the expert on the similarity of a small number of cohort participants. DIVA uses semi-supervised subspace clustering and subspace construction to identify sets of variables – subspaces – that promote participant similarity with respect to the outcome and with respect to the expert inputs, and then discovers subpopulations with increased outcome prevalence in those subspaces (DIVA module "DRESS"). DIVA uses visual analytics techniques to assist the expert in juxtaposing, filtering and inspecting the characteristics of these subpopulations (web-based DIVA module "D-INSPECTOR"). If the expert has access to a second cohort on a comparable population, DIVA aligns the cohort used for discovery to this second cohort, and then checks whether the subpopulations found in the original cohort are also present in the second one (DIVA module "VALIDATOR").

Results: We applied DIVA to the third wave (SHIP-2) of the SHIP-CORE cohort of the Study of Health in Pomerania (Völzke et al., 2011) for the liver disorder "hepatic steatosis", and on the first wave (TREND-0) of the SHIP-TREND cohort of the same study for the thyroid gland disorder "goitre". We found that most of the subpopulations extracted automatically, and subsequently ranked and filtered by the modules of DIVA, had significantly higher disease prevalence than the general population. We varied the amount of inputs needed from the expert to drive the subpopulation extraction process and found that a very small amount of information, namely the outcome of as few as 4 cohort participants, is sufficient for the identification of several relevant variables and subpopulations. We used a subset of TREND-0 for the validation on goitre and the complete TREND-0 for the validation on hepatic steatosis and found that the significant difference in prevalence for the identified subpopulation also holds in the validation data.

Conclusions: We have shown that DIVA discovers subpopulations and variables of importance with respect to an outcome, while requiring a very small amount of expert inputs. Each combination of variables and each subpopulation corresponds to a hypothesis, the validation of which would have required substantial

* Corresponding author.
E-mail addresses: tommy.hielscher@ovgu.de (T. Hielscher), uli.niemann@ovgu.de (U. Niemann), bernhard@isg.cs.uni-magdeburg.de (B. Preim), voelzke@uni-greifswald.de (H. Völzke), till.ittermann@uni-greifswald.de (T. Ittermann), myra@ovgu.de (M. Spiliopoulou).

https://doi.org/10.1016/j.eswa.2018.07.003

human effort. Thus, DIVA allows for a more effective exploitation of population-based data, not fully automated but driven by the expert and without the need for technical parameter tuning.

A shortcoming of DIVA design is the demand of a specific type of expert inputs, namely "constraints" on the similarity of pairs of participants. Currently, we generate the constraints with a naive utility that is based on random sampling, but we work on the development of an interactive algorithm that would allow the epidemiology expert to inspect a small choice of study participant and give statements on their similarity.

The present version of DIVA considers a single wave of the cohort data, ignoring the evolution of the population during the horizon of the study. Hence, subspace and subpopulation discovery do not take account of changes in the importance of variables. We currently work on the incorporation of algorithms that derive additional variables from the longitudinal data and use them in the Discovery module.

1. Introduction

Researchers in epidemiology collect population-based cross-sectional and longitudinal cohort data, from which they strive to derive insights on pathogenesis, disease pathways and responses to different kinds of treatments (Preim et al., 2016). Similarly to randomized clinical trials, a study on a population-based cohort involves an in/exclusion protocol and a carefully specified set of variables, whose impact on the outcome (e.g. on a disease) is to be investigated. Unlike randomized clinical trials, for which a cohort is prospectively recruited, studies on a population-based cohort are retrospective: the cohort has already been recruited with a protocol that typically involves a substantially larger number of variables. For example, the original protocol of the Study of Health in Pomerania encompassed 8854 variables (Völzke et al., 2011): analysis in such a high-dimensional space is prone to the curse of dimensionality, hence methods for focussed analysis in subspaces are necessary. In this study, we propose DIVA, an intelligent system with which the epidemiologist can semi-automatically, with minimal interaction, identify subsets of variables with potential relevance to a given outcome, can study automatically derived subpopulations that are described by these variables and exhibit considerably higher or lower prevalence of the outcome, and can juxtapose the significance of the prevalence difference in a validation cohort.

The task of supporting epidemiology experts with intelligent IT is being intensively investigated for years. Thew et al. (2009) elaborate on instruments with which epidemiologists can express and share domain knowledge among themselves. Such instruments are designed for hypothesis refinement. However, hypothesis generation, which comes before refinement, calls for decision with respect to (w.r.t.) the selection of the variables to be taken into account.

The selection of a small number of variables for hypothesis formulation from a huge set of variables has been studied by Zhang, Gotz, and Perer (2014) in the context of "cohort specification" from Electronic Health Records. Their system CAVA contains an interactive mechanism with which a clinical expert can select variables manually and study their impact on the outcome, as well as modules for automated management of the data in databases and for machine learning. Further systems in this category include SeekAView (Krause, Dasgupta, Fekete, & Bertini, 2016a), INFUSE (Krause, Perer, & Bertini, 2014) and PROSPECTOR (Krause, Perer, & Ng, 2016b), all of which include utilities for interactive variable selection before machine learning. However, the manual selection among hundreds or thousands of variables seems rather restrictive, since it enforces the expert to concentrate on the variables whose impact on the outcome is known or expected.

In this study, we propose a semi-supervised, self-tunable intelligent system that automates the construction of sets of variables potentially worth exploring, discovers subcohorts characterized by these variables, assists the expert in inspecting them, and validates them automatically in an independent cohort, if any is available. Our system DIVA consists of following modules:

- *Discovery* module: Our core algorithm DRESS + discovers subpopulations by exploiting little background knowledge in the form of pairwise constraints that contain knowledge about the similarity between study participants. Thus, the algorithm avoids the necessity of large quantities of labeled data by utilizing knowledge that can be derived from a limited set of labels or provided by a medical expert.
- *Inspection* module: Our interactive web application D-Inspector provides means to analyze the discovered subpopulations, i.e., juxtapose multiple subpopulations, study the distribution of corresponding variables w.r.t. the medical outcome, and query the set of subpopulations by custom filtering and sorting functionalities.
- *Validation* module: Our Validator checks to what extent the clusters and subspaces found by DRESS + can be reproduced in an independent cohort.

The paper is organized as follows. In the next two sections we discuss first related work and then basic underpinnings of the intelligent, semi-supervised technologies we use. In Section 4 we describe the three components of our approach. In Section 5 we present our results for the disorders fatty liver (hepatic steatosis) and goiter, using data from two cohorts of the Study of Health in Pomerania (SHIP) (Völzke et al., 2011). We close the paper with a discussion and outlook in Section 6.

2. Related work

Relevant literature for our work encompasses advances on intelligent systems for the support of the human expert in the medical domain, as well as advances on the functionalities covered by the modules of our system DIVA. We discuss them hereafter.

2.1. Interactive intelligent systems for cohort analysis

Without providing an integrated workflow to validate and inspect the findings, medical researchers remain skeptical towards the machine learning methods. Cummins (2012) describes that much criticism of the medical community on data mining models is due to the contrast between the sequential process of traditional medical research and the iterative and interactive approach of KDD procedures. To overcome this criticism, scholars should consider (i) involving domain experts into the model generation, (ii) assessing the model's quality by applying it on unseen data, and (iii) calibrating the model for different target populations (Cummins, 2012). Most of the systems described hereafter focus on expert involvement in requirement (i), while quality assessment is incorporated