Expert Systems with Applications 41 (2014) 5405–5415

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis

Uli Niemann[a], Henry Völzke[b], Jens-Peter Kühn[c], Myra Spiliopoulou[a,*]

[a] Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany
[b] Institute for Community Medicine, Ernst-Moritz-Arndt University Greifswald, Walther-Rathenau-Straße 48, D-17475 Greifswald, Germany
[c] Institute for Diagnostic Radiology and Neuroradiology, Ernst-Moritz-Arndt University Greifswald, Sauerbruchstraße, D-17487 Greifswald, Germany

ARTICLE INFO

Keywords:
Medical data mining
Classification rules
Subpopulation mining
Interactive data mining
Longitudinal epidemiological studies
Hepatic steatosis

ABSTRACT

Personalized medicine requires the analysis of epidemiological data for the identification of subgroups sharing some risk factors and exhibiting dedicated outcome risks. We investigate the potential of data mining methods for the analysis of subgroups of cohort participants on hepatic steatosis. We propose a workflow for data preparation and mining on epidemiological data and we present InteractiveRuleMiner, an interactive tool for the inspection of rules in each subpopulation, including functionalities for the juxtaposition of labeled individuals and unlabeled ones. We report on our insights on specific subpopulations that have been discovered in a data-driven rather than hypothesis-driven way.

1. Introduction

Medical research on epidemiological data aims to identify risk factors for diseases and to contribute thus to prevention and diagnosis (Preim et al., 2014). Epidemiological data come from *population-based* studies with a large number of randomly selected participants (the *cohorts*), for whom several variables are recorded; these may include sociodemographics, results of medical tests (e.g. on blood samples) and, recently, also magnetic resonance imaging (MRI) of body parts. Epidemiological research is commonly *hypothesis-driven*: scholars formulate a hypothesis on how a behaviour (e.g. alcohol consumption), a chronic disease (e.g. diabetes), a genetic predisposition or other factor may affect the prevalence of a certain outcome (e.g. cirrhosis, fatty liver); they then perform a careful selection of cohort participants with and without the outcome, on which the role of the candidate determinant is investigated with statistical analysis.

With the proliferation of medical engineering technology, an enormous number of variables, including biomarkers, can be recorded in population-based studies. Formulating hypotheses on each and every variable of those truly *Big epidemiological Data* is impractical, so that *data-driven* analysis of epidemiological data, namely epidemiological data mining, is now gaining momentum. Its purpose is to identify factors that are potentially associated with an outcome, so that hypothesis-driven analysis concentrates on them. This trend is further strengthen by the demands of personalized medicine (Hingorani et al., 2013), which requires the detection of previously unspecified subpopulations of patients that share common determinants (i.e. factors associated with an outcome).

In this study, we propose a mining workflow and an interactive tool for the discovery of potential determinants and corresponding value intervals that are associated with the multifactorial disorder hepatic steatosis (aka: fatty liver). Our emphasis is on highlighting fatty liver determinants that are not characteristic of the whole cohort but of cohort subgroups, e.g. female participants aged over 50.

Data mining methods are used widely on *clinical data* for diagnostic and therapeutic purposes, and there are also comparative studies on the performance of different mining algorithms for a specific clinical task, see e.g. (Pombo, Araújo, & Viana, 2014). However, clinical data mining analyzes data on patients, while epidemiology cohorts consist of participants with and without the outcome. Medical research on epidemiological data delivers the diagnostic indices that are later used for clinical diagnosis, e.g. the “fatty liver index” proposed by Bedogni et al. (2006). Hence, although our workflow encompasses mining methods that have been used on clinical data, studies on clinical data mining do not provide evidence on the potential of these same methods for epidemiological data analysis tasks, as we study here.

Mining methods on epidemiological data are rather rarely used. Multiple regression is often the approach of preference (Friedman & Meulman, 2003), but linear models, Cox regression and Poisson regression have also been used - notably on the epidemiological

* Corresponding author. Tel.: +49 3916758965.
E-mail address: myra@iti.cs.uni-magdeburg.de (M. Spiliopoulou).

http://dx.doi.org/10.1016/j.eswa.2014.02.040
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

data we analyze here (Baumeister et al., 2008; Haring et al., 2009; Lorenz et al., 2012; Völzke et al., 2013a). However, the use of regression in such studies is still mostly hypothesis-driven, e.g. on whether liver ultrasonography can predict mortality risk from elevated serum gamma-glutamyl transpeptidase levels (Haring et al., 2009). Our objective in this study is to demonstrate the potential of data-driven analysis for class separation in epidemiological data. Finding *new, previously unsuspected* determinants is not within the scope of our study. Rather, the potential of our approach is reflected on the quality of the learners, and, more importantly, on the data-driven identification of subpopulations that differ with respect to the class distribution, and on the data-driven discovery of associations that through independent, hypothesis-driven studies have earlier shown to exist.

We study our mining workflow on an multifactorial disorder, hepatic steatosis, using data from the first cohort of the "Study of Health in Pomerania" (SHIP); SHIP consists of population-based samples selected from Pomerania in Northeast Germany (Völzke et al., 2011). SHIP has already been extensively used for hypothesis-driven research on hepatic steatosis and lends itself excellently to the purposes of our analysis: we use the results of liver MRI recordings as target variable and a multitude of sociodemographic variables and medical tests for classification.

The contributions of our work are as follows. First, we propose a mining process for the classification of the participants of an epidemiological study with respect to a target outcome; we choose exemplarily the multifactorial disorder hepatic steatosis. Moreover, we propose an interactive tool, which we call InteractiveRuleMiner, with which a medical expert can drill into a derived model and investigate the properties of those subpopulations she considers interesting. Albeit mining workflows are often proposed for clinical data, mining methods for epidemiological data are rare and, in contrast to our method, they are hypothesis-driven. With our approach, a mining expert does not need to formulate hypotheses in advance, but can rather study the insights delivered by the models, identify subpopulations, drill-down on them and acquire further insights interactively.

The paper is organized as follows. In the next section, we discuss related work. In Section 3 we describe materials and methods for data preparation, population partitioning and classification. In Section 4 we report on the discovered models and important features for the different partitions. In Section 5 we present our tool "InteractiveRuleMiner". The last section concludes the paper with a discussion and an outlook towards learning disorder progression.

2. Related work

Medical decisions concerning the diagnosis of multifactorial diseases are based on clinical and epidemiological studies. The latter accommodate information on participants with and without the disorder and allow for discriminative model learning and, in the longitudinal design, for understanding the progress of a disorder (possibly towards a disease). There are several studies on the identification of factors (like obesity or alcohol consumption) and outcomes (like cardiovascular diseases) associated with hepatic steatosis. Findings on genetic and non-genetic factors include (Ittermann et al., 2012; Lau et al., 2010; Stickel et al., 2011); findings on associated outcomes include (Targher, Day, & Bonora, 2010 and Markus et al., 2013). However, these studies identify risk factors and/or associated outcomes that pertain to the whole population. Our study emanates from the necessity to identify such factors and outcomes for subpopulations and thus to stimulate personalized diagnosis and treatment, as expected in personalized medicine (Hingorani et al., 2013; Völzke et al., 2013b).

Classification on subpopulations is studied by Zhanga and Kodell (2013), albeit they analyze clinical data for diagnosis, while we analyze epidemiological data to identify variables associated with the outcome. Zhanga and Kodell point out that the complete population can be very heterogeneous, so that classifier performance on the whole dataset can be low. Therefore, they first train an ensemble of classifiers, then associate with each training instance the predictions made on it by each ensemble member, thus creating a new feature space where the variables are the predictions. They then perform hierarchical clustering on the instances, thus building three subpopulations: one where the prediction accuracy is high, one where it is intermediate and one where it is low. With this approach, Zhanga and Kodell (2013) split the original dataset into subpopulations that are easy or difficult to classify. The method seems appealing in general, but does not look promising in our case: we investigate a three-class problem with a very skewed distribution, so we already know that low accuracy is partially caused by the skew. Hence, we study the dataset exploratively *before* classification, to identify subpopulations that exhibit less skew, and exploratively *after* classification, to identify variables inside each subpopulation, which are associated to the outcome with high likelihood.

Pinheiro, Kuo, Thomo, and Barnett (2013) perform association rule discovery on patients with liver carcinoma. Pinheiro et al., 2013 point out that early detection of liver cancer may help reducing the five-year mortality rate (of currently 86%), but early detection is difficult, because in the onset of a liver carcinoma, the patient often observes no symptoms. The same authors further leverage the association rule miner FP-growth (Han, Pei, & Yin, 2000) to discover high-confidence association rules and high-confidence classification rules with respect to mortality in a liver cancer patients dataset. We also consider association rules promising for the analysis of medical data, because they are easy to compute and deliver results that are understandable by humans. Therefore, we also use association rules as baseline mining method, though for epidemiological data and for classification rather than mortality prediction. To use association rules for classification, we specify that the rule consequent should be the target variable (the rules are then called "classification rules"; we use this term hereafter).

Next to its advantages, association rule discovery (and classification rule discovery) has an inherent disadvantage: namely it generates large or even huge numbers of rules, among which the expert has to search for the truly interesting ones. Scholars have often proposed visualization as a remedy, and there is substantial research on comprehensible visual representations of large numbers of association rules. For example, Hahsler and Chelluboina (2011) group association rules' antecedents by their shared attributes to create a grid, where more important rules are displayed as circles; a circle's size and color reflects the rule's support and lift values, respectively. Further, their tool allows the user to zoom into interesting areas of the visualized set of rules.

Sekhavat and Hoeber (2013) stress that "...in spite of the advantages of previous works in visualizing association rules, the most common problem they encounter is their inability to handle a large collection of rules. In general, this results in occlusion and screen clutter problems due to the need to compress the visual representation into a single view." They propose SARV (Scalable Association Rule Visualisation), an interactive panel containing (i) a table-like grid view where rows represent rule antecedents, columns represent rule consequents, and a cell captures a single rule, colored (in grey-scale) according to the rules support value – this view is interactive, so that potentially interesting rules can be selected; (ii) a graph view for the visual exploration of the relationships between rules selected in the grid view; (iii) a textual view for displaying a rules support and confidence values (Sekhavat & Hoeber, 2013). The InteractiveRuleMiner of our approach also