International Journal of Approximate Reasoning 85 (2017) 159–177

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Dealing with under-reported variables: An information theoretic solution ☆

Konstantinos Sechidis [a], Matthew Sperrin [b], Emily S. Petherick [c], Mikel Luján [a], Gavin Brown [a,*]

[a] School of Computer Science, University of Manchester, M13 9PL, UK
[b] Centre for Health Informatics, Institute of Population Health, University of Manchester, M13 9GB, UK
[c] School of Sport, Exercise & Health Sciences, Loughborough University, LE11 3TU, UK

ARTICLE INFO

Article history:
Received 30 November 2016
Received in revised form 21 March 2017
Accepted 3 April 2017
Available online 7 April 2017

Keywords:
Under-reporting
Misclassification bias
Missing data
Mutual information

ABSTRACT

Under-reporting occurs in survey data when there is a reason for participants to give a false negative response to a question, e.g. maternal smoking in epidemiological studies. Failing to correct this misreporting introduces biases and it may lead to misinformed decision making. Our work provides methods of correcting for this bias, by reinterpreting it as a missing data problem, and particularly learning from positive and unlabelled data. Focusing on information theoretic approaches we have three key contributions: (1) we provide a method to perform valid independence tests with known power by incorporating prior knowledge over misreporting; (2) we derive corrections for point/interval estimates of the mutual information that capture both relevance and redundancy; and finally, (3) we derive different ways for ranking under-reported risk factors. Furthermore, we show how to use our results in real-world problems and machine learning tasks.

1. Introduction

Under-reporting usually occurs in survey data, and it refers to respondents that under-report the answer to a question, for example due to a perceived social stigma [35]. A famous example is maternal smoking during pregnancy, which is a key risk factor for adverse offspring outcomes including preterm birth and low birth weight (LBW). Like many health behaviours, accurate measurement of smoking habits can be difficult and expensive during pregnancy. For that reason, many studies use *self-reported* data, e.g. Wright et al. [37]. Given that most smokers know their habit to be harmful, both to themselves and their unborn child, there are strong motivations for women to under-report or deny their smoking status [10]. As such, the frequency of smokers in a sample is expected to be significantly lower than would be expected according to expert knowledge, derived for example from blood test result. Gorber et al. [15] presented a comprehensive analysis of the literature and compared the prevalence estimates of smoking based on self-reported data against the prevalence estimates based on directly measured smoking biomarkers. According to this analysis, self-reported smoking is generally under-reported in such a way that the true smoking figures may be underestimated by up to 47%.

☆ This paper is part of the Virtual special issue on the Eighth International Conference on Probabilistic Graphical Models, Edited by Giorgio Corani, Alessandro Antonucci, Cassio De Campos.

* Corresponding author.
E-mail address: Gavin.Brown@manchester.ac.uk (G. Brown).

http://dx.doi.org/10.1016/j.ijar.2017.04.002

Estimating the association between an *under-reported* (UR) variable and another will be biased in a manner that is specific to the degree/pattern of UR. Thus, any policy decisions made on the basis of such a biased result will be questionable. For example, government policies on tobacco control, e.g. [1], maybe ill-formed if they do not take into account UR. Maternal smoking and alcohol consumption are our focus for this paper, but there are many important health applications where *corrections* for UR are needed – e.g. HIV prevalence [38].

One method to correct UR bias is to spend time and resources to manually identify individuals that are likely to have misreported, and ignore/correct their testimony, i.e. identify smokers by performing cotinine blood tests. Unfortunately, in many applications it is impossible to completely correct the misreported cases. For example, Gorber et al. [15] present some possible flaws of the biochemical markers that identify smoking. As an alternative to this, authors in medical statistics treat it as a problem of *misclassification bias*, and combine data with a prior belief of the pattern of misclassification. They use this prior knowledge to derive corrected estimators for the log-odds ratio [7,12], and the relative-risk [27], or to suggest ways for performing tests of independence [5].

These solutions suffer from a number of weaknesses, which are addressed in this paper. For testing independence, they do not control both types of error (false positive/false negative). For estimating effect sizes, the suggested solutions are naturally only applied to estimate the correlation between a binary UR variable and a binary target variable – multi-class target variables with more than two categories are handled via a one-vs-one or one-vs-all strategy. Furthermore, ranking of variables in relation to a target – a common need in feature selection and other machine learning tasks – is not straightforward. Our strategy to overcome these limitations is to provide an information theoretic insight for the under-reporting problem, by disentangling three intimately related activities: *testing*, *estimation*, and *ranking* of features. Our main goal is to derive corrected estimators for the *mutual information* (MI), a measure of effect size widely used in machine learning applications with several interesting properties [4].

To achieve our goal we reinterpret the challenge not as dealing with misclassification and biased data, but as a problem of *learning from missing data*, and particularly learning from positive and unlabelled data [13]. By this interpretation, we present solutions using a graphical representation called *missingness* or *m*-graph [23], which is a tool to naturally incorporate a prior belief over the misreporting at the population (or appropriate sub-demographic) level. Furthermore, with our work, we show how to correct MI for under-reporting by examining independence properties observable via the *m*-graph representation.

In this paper, we present the following novel contributions[1] in relation to UR variables:

- Testing: **Section 3** suggests a way for testing independence between an UR feature and the class. Using our test and a derived correction factor, we can control both false positive/negative errors.
- Estimation: **Section 4** derives corrected estimators of MI terms that occur in UR scenarios. Section 4.1 presents estimators that capture the *relevance* between an UR variable and an arbitrary categorical variable, while Section 4.2 presents an estimator that captures the *redundancy* between two UR variables. Furthermore, we provide interval estimates where possible.

Section 5 presents two different methods for information theoretic feature ranking in the presence of UR variables, by using our suggested estimators. Section 5.1 presents Corrected-MIM, a univariate approach that provides rankings and captures only the relevance, while Section 5.2 presents Corrected-mRMR, a multivariate method that captures both the relevance and redundancy between the features. All the above contributions are novel, with the exceptions of Sections 4.1 and 5.1, which have been published in a conference paper [33]. Furthermore, we provide further experimental results in two applications. Firstly, **Section 6** derives rankings of risk factors related to low birth weight infants using a case study of 13,776 births in northern England, where we demonstrate some significant false conclusions that might be drawn when ranking variables without the correction factors. Finally, **Section 7** presents a machine learning application where we derive rankings of features when training/test distributions differ.

2. Background material

To the best of our knowledge, our work is the first that tackles the problem of estimating MI in under-reporting scenarios. In classic statistics there are some works that estimate other types of effect sizes (i.e. odds/risk ratios, limited to binary data) and we review them in Section 2.1. Section 2.2 shows how the under-reported can be phrased as a missing data problem. Finally, Sections 2.3, 2.4 and 2.5 give the background on testing, estimation and ranking using information theoretic measures.

2.1. Under-reporting as a misclassification bias problem

We assume that we have two random variables X and Y, representing a scenario where X is likely to be UR. In this case, we cannot observe the true value of X, but instead receive observations from a proxy variable $\widetilde{X}$. In the notation below we use lower case letters $(y, x, \widetilde{x})$ to denote a realisation from these variables. In our example of smoking during pregnancy,

[1] The software related to this paper will be available at github: https://github.com/sechidis.