MEDINFO 2013
C.U. Lehmann et al. (Eds.)

doi:10.3233/978-1-61499-289-9-922

Impact of variations in Anonymous Record Linkage on Weight Distribution and Classification

Daniel Nasseh, Jürgen Stausberg

IBE, Ludwig-Maximilians-Universität München, Germany

Abstract and Objective

Anonymous or privacy preserving record linkage is the term for systems allowing the linkage of data from different sources while maintaining an individual's anonymity. This work displays the impact of variations in the process of generating weights in a probabilistic record linkage system on different datasets, the resulting set of weights of candidate pairs and consequently on the final classification process. Furthermore, the results give insight into general problems of current unsupervised classification methods.

Keywords:
medical record linkage, privacy of patient data, classification.

Introduction

An example displaying the need of anonymous record linkage is the ongoing study of family-based cancer of colon in Germany [1]. In this case, anonymous record linkage is used to match patients recently diagnosed with having colon-cancer and their relatives to patient data of the Munich Cancer Registry. Preliminary work for this study has brought up observations pinpointing that even small variations in a probabilistic record linkage environment can severely alter the distribution of matching weights and therefore influence classification based on unsupervised learning. Exemplary, some of these variations have been selected in order to observe how the differences in weight generation can influence the determination of an appropriate cutoff-point which is used to partition record pairs into classes of links or non-links.

Methods

A configurable probabilistic record linkage system, written in java has been set up for testing. The system is based on the widely used algorithm of Felligi and Sunter [2]. The selected configurations vary in the way of blocking and in the addition of a post processing technique which in this work is referred to as multi-link-cleaning (MLC). This results in three different sets of weights for each of two different test sets consisting of a publicly available as well as an artificially generated test set. Blocking is a way to limit the calculation of weights, and therefore decreases computational resources to record pairs which exclusively agree in specific blocking variables. The two different ways of blocking used in this work differentiate in storing calculated weights for the different blocking variables uniquely or not. MLC is meant to remove all links which include a record which has already been part of another link with a higher weight. As a publicly available test set the relatively small 'Census' dataset has been chosen [3]. In addition, in order to obtain data which is more similar to the expected data in the study of colon cancer, artificial data based on attribute occurrences in different publicly available German datasets has been created.

Results

The six resulting sets of weights were visualized as histograms. Unsupervised classification techniques like support vector machines search for the largest margin within a set of weights and consequently splitting the set at the given position into links and non-links. Such a position can often visually be spotted as a deep notch or broad gap within the histogram of weights. Unfortunately, the largest notches or gaps are not always the best possible positions for a cutoff value. In the histograms, especially in the case of the publicly available dataset, it is not possible to spot a valid classification position without ambiguity due to the occurrence of many different deep notches and gaps. In the histograms for both the public as well as the artificial datasets one can recognize the similarity between the histograms for the three different configurations. Still, the impact based on relatively small changes is visible and can lead to different positions for the cut-off value.

Conclusion

The classification process depending on specific record linkage systems can be ambiguous. Therefore it appears to be doubtful to base the decision for positioning a classifying cutoff value on only one configuration if no further manual validation is possible. Consequently it appears that commonly used unsupervised classification systems can be unreliable. Therefore we suggest further investigations in the field of supervised learning classification which does not rely on the set of weights but on trainings data instead.

References

[1] Mansmann U, Stausberg J, Engel J, Heussner P, Birkner B, Maar C. Familien schützen und stärken – Umgang mit familiärem Darmkrebs. Gastroenterologe 2012: 161-162.

[2] Felligi I, Sunter A. A theory of Record Linkage. American Statistical Association Journal 1969: 1183-1220.

[3] Cohen W, Ravikumar P, Fienberg S, Kathryn R. Secondstring.sourceforge.net [Internet]. [cited 2012 Sep 19]. Available from: http://secondstring.cvs.sourceforge.net/viewvc/secondstring/secondstring/.