A Supervised Gradient-Based Learning Algorithm for Optimized Entity Resolution

Orion F. Reyes-Galaviz, Witold Pedrycz, Ziyue He, Nick J. Pizzi

PII: S0169-023X(16)30203-8
DOI: https://doi.org/10.1016/j.datak.2017.10.004
Reference: DATAK1621

To appear in: *Data & Knowledge Engineering*

Received date: 14 September 2016
Revised date: 23 August 2017
Accepted date: 14 October 2017

Cite this article as: Orion F. Reyes-Galaviz, Witold Pedrycz, Ziyue He and Nick J. Pizzi, A Supervised Gradient-Based Learning Algorithm for Optimized Entity Resolution, *Data & Knowledge Engineering*, https://doi.org/10.1016/j.datak.2017.10.004

A Supervised Gradient-Based Learning Algorithm for Optimized Entity Resolution

Orion F. Reyes-Galaviz[1*], Witold Pedrycz[1,2,3], Ziyue He[1], and Nick J. Pizzi[4]
[1]Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4, AB Canada
[2]Department of Electrical & Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, 21589, Saudi Arabia,
[3]Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[4]InfoMagnetics Technologies Corporation, Winnipeg R3C 3Z5, MB Canada
reyesgal@ualberta.ca
wpedrycz@ualberta.ca
ziyue8@ualberta.ca
pizzi@imt.ca

*Corresponding author

Abstract
The task of probabilistic record linkage is to find and link records that refer to the same entity across several disparate data sources. The accurate linking of records (entity resolution) is an important task for the healthcare industry, government, law enforcement, and the private sector, for obvious reasons. However, finding exact matches of an entity can be challenging due to records with typographical, phonetical or other types of errors (noise) found across real-world data sources. Over the years, many comparison functions have been developed to relate pairs of records and produce a similarity score. With a pair of predefined thresholds, one may decide if records pairs match, do not match, or if they require further clerical review. Nevertheless, finding appropriate comparison functions, identity descriptors (fields), threshold values, and efficient classifiers remains a challenging task. In this study, we propose a supervised gradient-based learning model that can adjust its structure and parameters based on matching scores coming from many comparison functions (and applied to many fields), to efficiently classify the records. The design of this structure is transparent, and can potentially allow us to locate which comparison functions and fields are more significant to correctly link the records. To train this structure, we propose a novel performance index that can help learn how to separate matched from non-matched records. Results completed with the use of synthetic datasets affected by different levels of noise show the effectiveness of the algorithm, which can significantly reduce the number of false positives, false negatives, and the number of records selected for review.

Keywords: Record linkage, entity resolution, field selection, comparison functions, clerical review threshold, autolink threshold, gradient-descent, decision model.

1. Introduction

The main goal of record linkage (entity resolution) is to identify, match, and merge records, across several disparate data sources that belong to the same entity [4], [8]. An entity is a core corporate object such as a person (patient, client, traveler, convict, tax payer, scholar, etc.), publication, business, consumer product, and alike. The record linkage challenge arises in many different areas, including: hospitals trying to track records of a single person across other hospitals or from several visits; companies that want to merge their datasets with other

*