PROJECT PROSPECTUS - UDACITY MACHINE LEARNING NANODEGREE

Record Linkage Using Noisy Demographic Data

A Machine Learning Approach

Submitted by: Chad Acklin
1/30/2018

DOMAIN BACKGROUND:

In real-world applications, data often is stored in different data sources with different architecture, data elements, and unique data quality considerations. It is often necessary to integrate these disparate data sources to gain a more complete picture of a particular “entity” or “person”. Matching these records, though, is often a challenging task because of the lack of shared identifiers like Social Security Number of other national ID. Systematically matching a record from “John Doe” to a record from another system for “Jonathan Do” is a common yet challenging data task. Such a task is critical in a variety of domains including healthcare[1], business[2], and counter-terrorism.[3]

This task has been noted and addressed in research literature for many years. Scholars as early as Dunn in 1946 have noted the needs for (among other things) a national system to reliably link records for birth and death to create what he refers to as the “Book of Life.” [4]

Approaches to this problem generally fall into two categories: deterministic (where the records must match completely on one or more data elements) and probabilistic (in which various weights are considered to in regard to agreement of available variables) .[5] Recent literature has focused on automated statistical models for record matching including Baysian and Neural Network methods. [6]

[1] Golden, Cordell, and National Center for Health Statistics (U.S.). *Linkage of Nchs Population Health Surveys to Administrative Records from Social Security Administration and Centers for Medicare & Medicaid Services, Programs and Collection Procedures*. Vital and Health Statistics, Series 1, Programs and Collection Procedure, No. 58. Hyattsville, Maryland: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2015.

[2] H. Kopcke, A. Thor and E. Rahm, "Learning-Based Approaches for Matching Web Data Entities," in *IEEE Internet Computing*, vol. 14, no. 4, pp. 23-31, July-Aug. 2010. doi: 10.1109/MIC.2010.58

[3] Gomatam, Shanti, and Michael D Larsen. "Record Linkage and Counterterrorism." *Chance* 17, no. 1 (2012): 25-29. doi:10.1080/09332480.2004.10554883.

[4] *Dunn, Halbert L. 1946. “Record linkage.” American Journal of Public Health and the Nations Health 36(12):1412–1416.*

[5] Tromp, Miranda, Anita C Ravelli, Gouke J Bonsel, Arie Hasman, and Johannes B Reitsma. "Results from Simulated Data Sets: Probabilistic Record Linkage Outperforms Deterministic Record Linkage." *Journal of Clinical Epidemiology* 64, no. 5 (2011): 565-72. doi:10.1016/j.jclinepi.2010.05.008.

[6] Wilson, D. Randall, D. Randall (July 31 – August 5, 2011). *Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage* (PDF). Proceedings of International Joint Conference on Neural Networks. San Jose, California, USA.

The author's specific interest in this project grows out of a real world problem encountered in industry in which an existing database of customers needs to be integrated with data from OCR scans of paper forms. Due to privacy concerns, the forms do not have identifiers like Social Security Number that would make the integration task much easier. A solution is sought to reliably connect the information from the paper form into the existing database using only the available demographics collected.

PROBLEM STATEMENT:

This study will classify two datasets based only on demographic information where noise is introduced into the source data in the form of OCR mistakes, phonetic misspellings, number transpositions, and missing data. Data from one data source may or may not have a "match" in the related dataset. Duplicate matches are also possibilities.

PROPOSED SOLUTION:

This study will evaluate the efficacy of various machine learning techniques to link records from two data sources. The author intends to explore the efficacy of Linear, Tree-based, Ensembled, and Multi-layered Perceptron approaches to the problem. The data will be paired between sources and classified as a match or non-match.

DATASET:

Due to privacy concerns, it is difficult to obtain publically available examples of real-world data. Because of this, the author will be using synthetically generated data. The data will be generated using the Freely Extensible Biomedical Record Linkage Project.[7] This project provides a data generator that creates records and randomly introduces noise that would be expected from OCR errors, keying errors, and phonetic misspellings.

The data for this task will consist of 30,000 original records with 15,000 matching records. I have generated the data using the FEBRL package and prepared for the machine learning task.

Comparing every original to every potential match becomes impractical as the size of the recordsets grow. For example, in this particular problem, the Cartesian join of the data would result in 450M possibilities (30,000 * 15,000).

In order to limit the number of comparisons required, a method of blocking is employed.[8] While the data elements are not all equal, they usually share some data in common. In order to

[7] Christen, Peter, Churches, Tim and Markus Hegland "Febrl - A Parallel Open Source Data Linkage System: Proceedings of the 8th Pacific-Asia Conference", PAKDD 2004, Sydney,
Australia, May 26-28, 2004. Pages 638 - 647. *Springer Lecture Notes in Artificial Intelligence, Volume 3056.*

[8] Steorts R.C., Ventura S.L., Sadinle M., Fienberg S.E. (2014) A Comparison of Blocking Methods for Record Linkage. In: Domingo-Ferrer J. (eds) Privacy in Statistical Databases. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham

generate the dataset for a machine learning model, records are “blocked” based upon the following criteria:

1. Phonetic Matches:
    a. First Name
    b. Last Name
    c. City
2. Exact Matches
    a. Zip Code
    b. Phone Number
    c. First 2 letters of the last and first names

Each original record is compared to the blocked potential matching record and then features are engineered to determine the string similarity between the two datasets. A combination of exact matching and edit distance (Levenshtein distance) of each pair of data elements are employed. I calculated edit distance using the jellyfish python package. The specific metric used is the Damerau-Levenshtein Distance, which measures the number of edits & transpositions required for two strings to be equal.[9] I scaled and inverted the edit distances so that a value of 1 represents a perfect match and values of <=0 are completely different.

The data preparation steps and sample data have been attached to this proposal.

BENCHMARK:

The resultant model from the machine learning approaches tested will be compared to a deterministic matching model based on the demographics available. Because it is found in the literature[10], I will use a simple implementation of a Gaussian Naive Bayes Model as a baseline.

EVALUATION METRICS:

Results from the classification task (match/non-match) will be evaluated on precision, recall, and f1-score. Success is defined as a model that obtains a better f1-score than the baseline model on the test set of data.

PROJECT DESIGN:

The following steps will be followed to build the record linkage model:

1. Generate data using FEBRL data generator. Two datasets will be generated. One for training and testing models and the other to validate results.
2. Matching records has been “blocked” in the manner discussed above in order to generate a smaller subset of data matches to evaluate.

[9] http://jellyfish.readthedocs.io/en/latest/comparison.html#damerau-levenshtein-distance

[10] Koudas, Nick, Sunita Sarawagi, and Divesh Srivastava. "Record linkage: similarity measures and algorithms." In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 802-803. ACM, 2006.

3. Feature Engineering. Features will be created based on exact matching, phonetic and "fuzzy" matching, and string difference measures such as Levenstein distance and others.
4. Exploratory data analysis. Data will be explored using statistical and visualization tools to begin identifying important variables and potential trends.
5. Split the data into training and testing sets. ⅓ of the data will be held out as a test set, using the remainder for training. Sci-kit learn's train_test_split will be used for this task.
6. Build Gaussian Naive Bayes benchmark model
7. Investigate the efficacy of different machine learning approaches. Data will be split into training/testing sets and fed into various machine learning models. The author intends to explore at least one model from each of the following categories:
    a. Generalized Linear Models
    b. Tree-based classifiers
    c. Ensemble classifiers
    d. Multi-Layered Perceptron
8. Choose one model and tune parameters for improved results. Feature importances will be considered in order to build a model with appropriate tradeoff between accuracy and scalability.
9. A successful outcome will be to improve upon the f1-score of the benchmark model on the test set.