Single-person household deduplication for Colombia CO

Daniel Alvarez, Data Assurance

October 31, 2019

Single-person household deduplication for Colombia CO

This document presents a Naive Bayes model to determine likelihood of individuals being duplicitously registered across datasets.

Presentation of the problem

A concern raised in the provision of humanitarian assistance in the context of a country with multiple humanitarian programs has been the possibility of beneficiaries receiving more than their entitled benefits across multiple programs. An example would be a beneficiary family successfully registering across multiple programs, using different forms of identification, in order to receive more than their entitled benefits. In this scenario, humanitarian providers want to be able to identify such beneficiaries and perform an adjudication of identity. The problem arises that identifying beneficiaries may be complicated by the use of different forms of identifying documentation and different identifying information set across different program registration processes.

In a context in which the majority of beneficiaries are single-person households (individuals), there is no individual key to all name and other identifying information pertaining to a unique person and there is no biometrical information across datasets, there exists the perennial challenge of identifying duplicitous individuals. Identifying uniqueness of an individual when information is either erroneously transcribed or different documentation is used for recording registration across data systems requires probabilistic methods for classifying an individual as the person it should be.

As a response to this challenge, what follows is the presentation of a Naive Bayes probabilistic model for individual level classification. Classification is interpreted as the identification of an individual as a function of the evidence that has been collected about the individual. In principle, no matter how much distinct evidence exists for a given individual, there should be some small probability for the set of evidence to be associated with a new individual appearing in a dataset.

Given that there are many individuals in the dataset, what follows is the mathematical proof of a multinomial Naive Bayes probabilistic model for identifying individuals across datasets. The question this answers is: can we identify new individuals appearing across datasets, given their recorded attributes, as being the same as existing individuals in the datasets?

Proof of mathematical concept

The mathematical concept starts from the premise of a new individual needing to be identified as the same as an existing individual in the dataset.

$P(y_i|X) = P(y_i) \prod_{i=1}^{n_d} P(x_{i,k}|y_i)$

where:

X: feature set of individual attributes; where $x_{i,1}$, $x_{i,2}$ and $x_{i,3}$ are the first, second and third individual attributes for a given individual.

y_i: individual identity (i.e. the classifier)

$P(y_i|X)$: posterior probability of an individual being identified as an existing individual given a set of individual attributes

$P(y_i)$: prior probability of an individual being of a given identity. This can be thought of as the probability of an individual identity being in the dataset.

$P(x_{i,k}|y_i)$: conditional probability of a given attribute, for a given individual, being attributed to a given identity. For example, the conditional probability of a given individual being female gender is for a given identity.

Note on the subscripts:

k: denotes a given attribute

i: denotes a given individual

n_d: denotes the number of unique individuals in the dataset

Once the parameters, $P(y_i)$ and $P(x_{i,k}|y_i)$ have been estimated from a training dataset, given a new individual, the output of the Naive Bayes classifier is:

$argmaxP(y, x_i...x_d) = argmax\big(P(y_i)\prod_{i=1}^{n_d} P(x_{i,k}|y_i)\big)$

Worked example using data:

Setup

In this worked example, the features were taken from the common columns across datasets in the `DATA SET (SCOPE-EKAA-PROGRES- CCD)` excel file provided by the Colombia CO. The features considered common were as follows (where the abbreviations in parenthesis are used for references in the results):

- Household Name
- Location (LOC)
- Address (AD)
- Document Type (DT)
- Document Num (DN)
- Last Name (LN)
- First Name (FN)
- Household Role (HR)
- Gender (G)
- Marital Status (MS)
- Date of Birth (DOB)
- Age (AG)
- Mobile Number (MN)
- Physical Disability Status (PD)
- Breastfeeding (BF)

For the purposes of demonstration, non-real data was used to populate 104 entries. These entries served to train the Naive Bayes learning algorithm. We call the latter dataset the *training set*. To test the learning algorithm, additional non-real data entries were generated based on an alteration of a single entry in the training set. The latter dataset is the *test set*. The test set is used to validate the performance of the Naive Bayes learning algorithm.

The Household Name code serves as the identity classifer of interest, y_i. The other features serve as the feature set, X. The prior, posterior and conditional probabilities are calculated from the data in the algorithmic process.

Results

The algorithm's performance is assessed by whether the predicted identity classifier (the Household Name) correctly matches the true identity classifier based on the set of features. The algorithm produces probabilities

for all potential identifier classifiers. When the predicted probability for the true identity classifier is the highest among the potential identifier classifiers, there is an identity match. That is, the predicted identity classifier will be equal to the true identity classifier.

The results below show the performance of the learning algorithm on a test set where each entry has a true identity classifier set equal to one of the entries from the training set. The test is to see how well the true identity classifier is predicted after learning on the training set. In the test set, small modifications are made to the entry from the training set corresponding to the true classifier in order to mimic potential transcription differences or different information provided across data sources in the real world.

The results are depicted in the following table below. The `x`'s indicate fields where modifications were made to the features from the training set to the test set. Blanks mean that no modifications were made to the features from the training set to the test set. The *ProbTrue* is the probability of a classifier being the true classifier given the feature set. The *ProbPred* is the probability of a classifier being the predicted classifier given the feature set. The *Match* field indicates whether there is an identity match.

Table 1: Results of Naive Bayes learning algorithm

ï..Match	Prob_True	Prob_Pred	LOCAD	DT	DN	LN	FN	HR	G	MS	DOBAG	MN	PD	BF
Yes	83.0%	83.0%												
Yes	44.0%	44.0%	x											
Yes	23.0%	23.0%	x	x										
No	7.0%	28.0%	x	x	x									
No	7.0%	28.0%	x	x	x	x								
No	0.9%	24.0%	x	x	x	x	x							
No	0.3%	29.0%	x	x	x	x	x	x						
No	0.4%	22.0%	x	x	x	x	x	x	x	x				
No	0.1%	27.0%	x	x	x	x	x	x	x	x	x			
No	0.1%	29.0%	x	x	x	x	x	x	x	x	x	x		
No	0.1%	28.0%	x	x	x	x	x	x	x	x	x	x	x	
No	0.1%	39.0%	x	x	x	x	x	x	x	x	x	x	x	x
No	0.3%	24.0%	x	x	x	x	x	x	x	x	x	x	x	x
Yes	80.0%	80.0%												x
Yes	80.0%	80.0%											x	x
Yes	33.0%	33.0%										x	x	x
Yes	31.0%	31.0%									x	x	x	x
Yes	21.0%	21.0%								x	x	x	x	x
Yes	20.0%	20.0%							x	x	x	x	x	x
No	11.0%	15.0%					x	x	x	x	x	x	x	x
No	0.9%	28.0%				x	x	x	x	x	x	x	x	x
No	0.9%	28.0%			x	x	x	x	x	x	x	x	x	x
No	0.4%	27.0%		x	x	x	x	x	x	x	x	x	x	x
No	0.4%	27.0%		x	x	x	x	x	x	x	x	x	x	x
Yes	32.0%	32.0%			x	x								
No	10.0%	43.0%		x	x	x								

Discussion of results

There are a few salient inferences that can be made from observing the results:

- The algorithm performs quite well given the limited amount of training data with identity matches in the most expected cases (i.e. cases where there is limited modifications to the features between data

sets).

- The algorithm performs poorly when there is significant modification to the features between datasets. Intuitively, the *ProbTrue* is lowest when most features are modified between datasets. This indicates that complete alteration or transcription differences across common features in the different data sources will constrain any ability to identity match.
- The algorithm finds a match in the case when just the document type (DT) and document number (DN) are modified and all other features are the same across data sources.
- The identity matches seem to indicate that modifications to some features, taken together, may not be critical for the matching analysis. These features, taken together, of less apparent relevance for matching are: Gender (G), Marital Status (MS), Date of Birth (DOB), Age (AG), Mobile Number (MN), Physical Disability Status (PD) and Breastfeeding (BF). It may be that these features do not impose as large of a conditional weight in the matching algorithm. Further analysis may be warranted to understand why.

Next steps

The following next steps would enhance the implementation of an identifier matching algorithm:

- Add more training data including more variation and keep performing testing to assess algorithm performance. The worked example just used 104 non-real data entries, which is much smaller than the real population of data.
- Modify the non-real datasets with real-world features, such as the true data structures corresponding to the common features used in the worked example. For example, the actual character length and structure of the document numbers. This will improve the learning ability of the algorithm on real-world data and allow for a better assessment of performance.
- Implement the Naive Bayes learning algorithm presented in this paper on a sample of real data across multiple data sources, if possible.