Methods of Detecting Home Language Shift in Canadian Census Data

Chris Choy, Kiefer Co, Matthew Fogel, Clarke Garrioch and Katie Martchenko
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
Email: kleung@cs.umanitoba.ca

***Abstract*—Canada is a nation composed of a highly diverse language population. This provides a unique opportunity to study the factors causing certain languages and language families to be lost over subsequent generations amongst allophones (people with a mother tongue other than English or French). This paper applies and compares the performance of Decision Tree Induction, Random Forest, and Categorical Naive Bayesian algorithm to census microdata to analyze the influence of various social and economic factors on the probability that allophones adopt official languages as their language spoken at home.**

***Keywords*-component; language cohorts; allophones; mother tongue; language persistence**

I. Introduction

Canada is a nation composed of a highly diverse language population. Immigration and migration has risen both in numbers and as culturally relevant components of modern communities, especially in diverse countries such as Canada.

This provides a unique opportunity to study the rates at which certain languages and language families are lost over subsequent generations as allophones (people with a mother tongue other than English or French) adopt English or French as their primary language. Certain factors such as sex, age, educational status and economic success may prove to be a key indicator of how quickly an individual adopts a language other than their mother tongue in everyday life.

A language shift occurs when an allophone adopts an official language as their primary language, i.e. language spoken at home. Several studies have aimed to measure language shift rates through linear regression on various cohorts of the population. Ultimately, it is impossible to ascertain precisely when a language shift occurs, so the insights offered by linear regression are limited in accuracy.

This paper proposes an application of various data mining algorithms and compares their accuracy and speed when used on census data, namely the Random Forest algorithm, the Decision Tree algorithm, and the Categorical Naive Bayesian algorithm.

Some challenges working with census microdata include the fact that the Public Use Microdata Files (PUMF) have been downsampled from the census population size. In order to perform an analysis that would take into account the population distribution, we needed to multiply records by a corresponding weight included in the PUMF datasets. Additionally, the PUMF datasets contain over 100 potential features, some of which contain largely invalid/unavailable data. As a result, some level of feature selection is required.

The Canadian census is conducted every five years. As a result, changes that occur between census periods may not be captured at the exact time of their occurrence.

II. Related Work

Over the past several years, researchers have come up with multiple approaches to analyze census microdata to determine the rates at which allophones express a language shift. Several authors has focused on identifying whether a shift towards official languages (English and French) have occurred by determining if the mother tongue is the same as the language spoken at home. This is typically accomplished by performing linear regression.

A. Linear Regression of Language Cohorts

1) Fictitious Cohorts: Patrick Sabourin and Alain Bélanger use the concept of a 'fictitious cohort', in which groups separated by age or time since immigration are compared across a single census. They define language persistence as the proportion of each cohort that has kept their mother tongue as the language most often spoken at home. The authors analyze language shift using linear regression and polynomial regression. They then construct a survival curve which determines the probability that each subsection of a cohort (such as a specific age group) will undergo a language shift. They make the assumption that the rate of language shift is constant among several censuses. [1]

One limitation to this method is that members of a cohort are defined in binary terms as having lost a language if they no longer speak it at home, while this process might occur gradually in real time. Additionally, certain language groups may experience different rates of language shift. [1]

2) Synthetic Cohorts: Marie T Mora, Daniel J Villa and Alberto Davila use an alternative method known as 'synthetic cohort' analysis on census data in the United States. Their paper aims to better understand the recent dynamic of language loss and intergenerational maintenance of Spanish in the U.S., and compare it to other non-English languages. In other words, exploring the retention or loss of Spanish and other non-English speakers in the U.S., particularly among foreign-born and U.S.-born children with immigrant parents. [2]

The technique for analysis, synthetic cohort analysis, is based on data drawn from 1980, 1990, and 2000 United States Censuses. It creates a temporal representation of a population, over ten-year intervals. The authors track the reported language use of individuals starting at ages 5-7 and ending at ages 15-17 across two United States census. This selection is due to children being able to speak at ages 5-7 and languages tending not to be lost after ages 15-17. [2]

This is in contrast to cross-sectional methods which use data from only one Census period to analyze language shifts. Combining this method with the synthetic cohort, the paper argues that the dynamic in language shift is better predicted, supported by what has been observed in the U.S. [2]

There are still some difficulties with this approach. For example, 1980 and 1990 samples could have emigrations before 1990 and 2000, and from this the "true" cohorts of the foreign-born may not be entirely reflected.[2]

3) Limitations of Linear Regression: As seen above, multiple approaches to analyzing language cohorts temporally run into limitations in how census data is collected. Sampling errors and poorly worded census questions make it difficult to capture whether emigration has occurred between censuses. Even within a census, it is tempting to view language shift within a cohort as binary when this process occurs gradually.

Populations are dynamic, and multiple categorical variables influence whether a language that is the mother tongue is spoken at home. A decision tree can reveal which categorical variables determine whether mother tongue is retained as the language spoken at home.

B. Decision Tree Performance in Other Areas

1) Decision Trees in Student Performance Prediction: Decision trees have also been compared and applied in other fields. Osmanbegovic and Suljic compared the performance of an implementation of a decision tree algorithm, Naive Bayes algorithm, and a Multilayer Perceptron algorithm in predicting student performance by the prediction accuracy, learning time, and error rate. [3]

Osmanbegovic and Suljic used 12 input features such as gender, GPA, and whether or not the student had scholarships, and outputted whether or not the student would pass or fail. The output could also be classified by letter grades, but due to the disparity in the amount of data for each class, was not used. [3]

They found that the Naive Bayes algorithm managed to outperform the C4.5 decision tree algorithm implementation, J48, in both prediction accuracy, and error rate. Additionally, it was found that the Naive Bayes algorithm and the decision tree algorithm created prediction models that were both accurate, and user-friendly enough for the stakeholders. [3]

2) Improvements: Similarly to Osmanbegovic and Suljic's data, some classes in census data are going to have differing amounts of data, and will affect prediction accuracy. Instead of changing the class groupings to get more equal representation, weights could be added to the training data to account for these differences. In the case of the Canadian Census data, these weights are already added to the data for use.

C. An Alternative Approach to Census Data Mining

In their paper, Klösgen and May took a different approach to mining census data. Klösgen and May used the United Kingdom's census data, which was only available in an aggregated form. The census data was available aggregated across various wards within the region, along with a detailed set of geographic layers. Thus, Klösgen and May decided to use those wards as the focus of their examination, and propose an application of SubgroupMiner, an advanced subgroup mining system. [4]

D. An Improved Decision Tree Algorithm

Hulten, Spencer, and Domingos' CVFDT algorithm improves on the VFDT decision tree learner by accounting for data changing over time. [5] Their CVFDT algorithm works by maintaining a decision tree with respect to a sliding window of data and grows an alternative subtree to replace an old one if it becomes out of date. This allows for it to learn a similar model to one from the VFDT algorithm but in constant time [5].

Applying the algorithm to census data seems like a natural fit, and can be used to improve the performance of any decision tree learners in use. The addition of the time aspect could also be used to improve the accuracy in cases where the data from only one period of time is used, instead of multiple.

III. Main Body

A. Data Collection

Data was collected from the 2016 Canadian census public use microdata file (PUMF), which contains around 930,421 (or 2.7% of the target population) individual, anonymized records, with 123 features. Additionally, there's an individual weight attached to each record, and 16 estimate weights for sampling variability.

1) Privacy: While the PUMF does give individual records, some of the data was aggregated to preserve confidentiality (e.g. categories being combined together), and some records had some of their variables changed to 'Not Available' for similar reasons. Furthermore, only the largest of the census metropolitan areas and provinces were covered.

2) Data Weighting and Sample Universe: Since the PUMF data is only a sample of the target population, each record includes an individual weight to indicate how much of the target population that the record represents. In addition to the weights, 31 of the features are drawn from the universe of family, household, and dwelling universes, and the remaining 92 features are from the individual universe.

3) Data Cleaning: Since some values in the PUMF data are marked as invalid or unavailable, some preprocessing on the data is required. Due to each record having an individual weight attached to it, rather than impute a value for the missing/invalid values, those rows were dropped instead.

In general, some features had a high rate of invalid data and were not used for any of the classifiers.

B. Methodology

Our analysis is performed in Python via Jupyter Notebooks. Packages utilized include **pandas** for data manipulation, **scikit-learn** for implementations of the classification algorithms in question, and **bokeh** and **tabulate** for visualization and formatting of our data.

1) Data Wrangling: Prior to the execution of any classifiers, the PUMF dataset is passed through a single data preprocessing pipeline. This pipeline achieves two broad objectives, namely pruning uninteresting data and selecting against features containing a high proportion of invalid data.

The feature selection process performed was a mix of manually intuited choices and insights from initial analysis using scikit-learn's decision tree classifier on unprocessed data. Features which were pruned were both uninformative with regards to our goal of detecting home language shifts and possessed clearly understood reasons for their behaviour. For example, MTNEN (mother tongue being English) was extremely negatively correlated with our shift metric, as English is the most spoken language in Canada and an unlikely home language to shift away from. Conversely, LWAEN (language at work being English) and LWAFR (language at work being French), had little to no influence on our shift metric, which was understood to be the consequence of work language rarely being a language other than English or French. Such features were omitted from the preprocessed version of our dataset.

The PUMF dataset also contained missing, invalid, or inapplicable values as specified by certain numeric codes (e.g. 88888888) as specified in its accompanying Study Documentation. Our initial preprocessing involved detecting the number of such values in each feature column as a proportion of the total number of entries. Features which possessed an unususable percentage greater than 20 were pruned so as to not obscure our analysis by training classifiers on presence of a data feature rather than its content.

Our initial data anaylsis also found promising improvements when dropping rows featuring English as a mother tongue. Removal of these rows lowered the unusable proportion of many features we intuited would provide useful insight, such as AGEIMM (age of immigration) which dropped from 88 to 61 percent unusable. Although AGEIMM in particular remained unsuitable for our analysis as only a fraction of respondents provded a usable value, we opted to remove English mother tongue rows from our dataset with the intention of providing more focused analysis on changes within Canada's non-official language communities, from which shifts to English as a home language were more common than the inverse.

As mentioned, data imputation was considered as a remedy for unsusable data, but opted against due to the structure of the PUMF dataset. The dataset is meant to balance presenting a representative subset of Canada's demographics while also anonymzing census results which, by Canadian law, are to be withheld from complete public disclosure for 92 years post-response. As such, each row in the PUMF dataset is a weighted representation of several individuals, listed under the attribute WEIGHT. Imputation of unusable values, even with the feature's weighted mean as a substitution, was reasoned to be unapplicable without rendering the dataset no longer representative of Canada's demographics.

2) Shift Detection: Our process for identifying language shift involves constructing a Boolean feature we call **languageShift**, computed from the MTNNO (mother tongue) and HLANO (home language) attributes.

In the PUMF dataset, both the MTNNO and HLANO features are categorical attributes encoded as a numeric value. Each feature's encodings are independent and cannot be readily compared, although there exists an intersection between the categories represented within each feature. Dictionaries for mapping MTNNO and HLANO categories to equivalency classes were manually defined using information outlined in the PUMF Study Documentation file. These classes include both whole languages such as "Arabic" as well as grouped categories such as "Austro-Asiatic languages".

As the MTNNO categories are less broad and appear to be a superset of the HLANO categories, some categories appear in MTNNO but not HLANO, such as "Uralic" or Korean". Rather than omitting rows containing these mother tongues, we opted to categorize them under the "All other languages" class.

The value of our languageShift attribute is then defined as the presence of inequivalent class mappings in the MTNNO and HLANO features within a data row. The reasoning behind this decision is that a difference between mother tongue and home language implies a shift having occurred at some point within an individual's lifetime, with the assumption made being that an individual's mother tongue must at one point have been their home language, possibly only in childhood or even outside of Canada.

If MTNNO can be mapped to HLANO, the languageShift value of a row is false. If no mapping can be found, the languageShift value will be true. As with other features in the dataset, the results of the languageShift column are to be interpreted as representative of several individuals. A row's weight is to be factored into any analysis on or prediction of the languageShift feature.

C. Decision Tree

Our data analysis notebooks load the preprocessed dataset and perform implementation-specific data manipulation prior to analysis. Our analysis pipeline involves manipulating and splitting the dataset, then training a classifier on the data. The classifier is examined to reveal insights on relevant features.

The languageShift and WEIGHT attributes are first separated from the dataset to serve as class labels and weights respectively. MTNNO and HLANO are dropped from the dataset as they are directly involved in the computation of languageShift, for which their effects on a classifier for would be uninteresting. The remaining attributes in the preprocessed dataset are then to be used as classification features.

The manipulated datasets are then split into train and test sets, with one quarter of the data reserved for testing and the remainder being allocated for training.

A decision tree is then trained to predict languageShift using the classification features and with respect to the data's weights. The tree is allowed to split on features until all leaves are pure, using gini to evaluate splits.

The precision and recall per languageShift class for the trained decison tree, tested on the testing dataset, are as follows:

	precision	recall	f1-score	support
False	0.89	0.88	0.89	75521
True	0.60	0.61	0.61	21434
Total			(accuracy) 0.82	96955

Feature importances are then extracted from the trained classifier. The top ten features obtained via this method are as follows:

Feature	**Importance**
GENSTAT (Generation status)	0.196041
VisMin (Visible minority status)	0.0601858
ETHDER (Ethnic origin)	0.0517905
POB (Place of birth)	0.0438852
PR (Province, current)	0.0398168
SHELCO (Shelter cost)	0.0390179
AGEGRP (Age)	0.0303643
POBF (Father's POB)	0.0285193
MrkInc (Market income)	0.0230586
ROOMS (# Rooms in dwelling)	0.0224302

The decision tree is then visualized using bokeh as seen in the following figure: *To be included.*

D. Random Forest

Our random forest analysis takes in data in a format identical to that of our decision tree analysis. Data is manipulated and split as described above, then used to train and evaluate a random forest classifier. The classifier in question uses twenty decision trees as estimators within its averaging process. A controlled random state is passed which acts as a seed to ensure repeatability of results.

The precision and recall per languageShift class for the trained random forest classifier, tested on the testing dataset, are as follows:

	precision	recall	f1-score	support
False	0.88	0.95	0.91	75523
True	0.76	0.55	0.64	21432
Total			(accuracy) 0.86	96955

Feature importances are then extracted from the trained classifier. The top ten features obtained via this method are as follows:

Feature	**Importance**
GENSTAT (Generation status)	0.063524
ETHDER (Ethnic origin)	0.0546316
POBF (Father's POB)	0.0473872
PR (Province, current)	0.0456554
POB (Place of birth)	0.0421344
POBM (Mother's POB)	0.0392552
DPGRSUM (Population group)	0.027472
SHELCO (Shelter cost)	0.0272865
PR5 (Province, 5 years ago)	0.0252248
VisMin (Visible minority status)	0.0242537

E. Naive Bayesian

An analysis on the Naive Bayesian algorithm will be included in the final version of the paper. Methodology used is similar to the random forest and decision tree analysis.

IV. Analytic Evaluation

V. Conclusion

Results will be placed here in the final version of the paper.

Acknowledgment

The authors would like to thank the staff at the Elizabeth Dafoe library for giving direction on locating census microdata, and Dr. Carson Leung of the Databases and Data Mining Laboratory at the University of Manitoba for support of this project.

REFERENCES

[1] P. Sabourin and A. Bélanger, "The dynamics of language shift in canada," *Population*, 2015.

[2] M. T. Mora, D. J. Villa, and A. Davila, "Language shift and maintenance among the children of immigrants in the u.s.: Evidence in the census for spanish speakers and other language minorities," *Spanish in Context*, 2006.

[3] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, 2012.

[4] W. Klösgen and M. May, "Census data mining - an application," 2012.

[5] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," 2001.