A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

Juan Sebastián Marín-Delgado

Tutor:
Francisco Zuluaga-Díaz

Oral Progress Presentation
April 8th, 2015

Agenda

1. Problem Statement

2. Variables of Interest

3. Data preprocessing

4. Algorithm Description

Problem Statement (1/3)

The idea before the segmentation of the US Hispanic households is to find a set of subgroups of the population, so that that the households within each subgroup are homogenous enough to find common characteristics or patterns.

The previous problem can be thought as a clustering problem as follows: let H be the set containing the Hispanic household sample, we want to find a set of subsets of H such that:

$$\cup_{i=1}^{k} C_i = H \quad (1)$$

$$\forall i = 1, \ldots, k \;\; C_i \neq \emptyset \quad (2)$$

$$\forall i \neq j, \text{for } i, j = 1, \ldots, k \; C_i \cap C_j = \emptyset \quad (3)$$

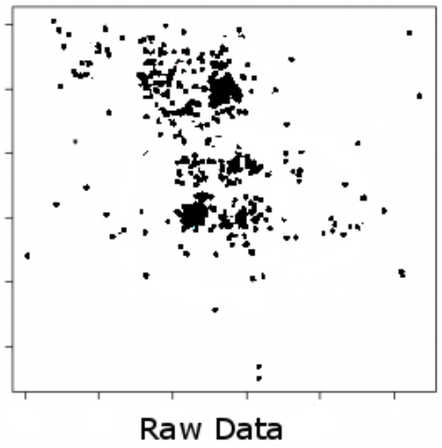
Problem Statement (2/3)

In addition, the set of clusters need to satisfy certain statistical properties to ensure homogeneity within the households of a given cluster and heterogeneity between the different clusters.

In order to generate the clusters fulfilling the previous conditions it becomes necessary to develop a clustering algorithm capable of generating valid solutions. It is also important to notice that the algorithm should take into account the the type of data describing the US Hispanic households.

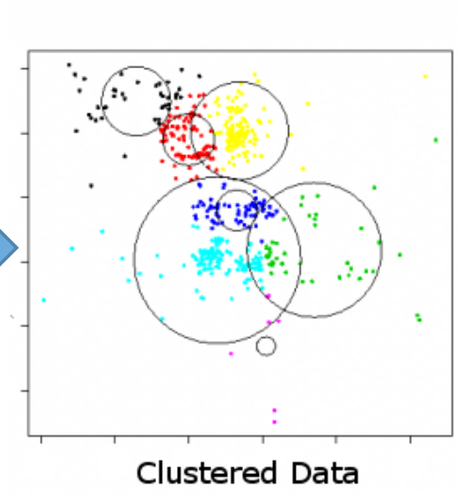
Problem Statement (3/3)

Numeric Data



Raw Data

Clustering algoritmm



Clustered Data

Categorical Data

Hair color → / Eye color ↓	Blonde	Black	Red	Brown
Green	2	2	0	1
Black	0	1	0	2
Brown	2	1	4	3
Blue	2	1	1	0

Clustering algoritmm

?

Advantages:
It is easy to define a distance function. For instance, the Euclidian distance seems to work pretty well for numeric data

Variables of Interest

For our problem we considered the following variables to be taken into account for the clustering analysis:

VARIABLE NAME	TYPE
STATE	Categorical
FES	Categorical
HOUSEHLAN	Categorical
MULTGEN	Categorical
RELCHILDREN	Categorical
OWNERSHIP	Categorical
HOUSEHINC	Numeric
AGE(5-)	Numeric
AGE(6-17)	Numeric
AGE(18-24)	Numeric
AGE(25-34)	Numeric
AGE(35-54)	Numeric
AGE(55-64)	Numeric
AGE(65+)	Numeric
ACCI(20-)	Numeric
ACCI(21-40)	Numeric
ACCI(41-60)	Numeric
ACCI(61-80)	Numeric
ACCI(81+)	Numeric
LHIGHS	Numeric
SOMEHS	Numeric
FINISHS	Numeric
SOMECL	Numeric
ASSORBACH	Numeric
MASORPHD	Numeric

Data Preprocessing

Preprocessing of data was executed as follows:

1. Subsetting the Census data by those who were categorised by the Census as Hispanic Households.

2. Extracting the variables of Interest

3. Cleaning the data

4. Splitting the data into Mexicans and Non Mexicans

Algorithm Description (1/4)

- The algorithm is based on the approach proposed by (Ahmad & Dey, 2007)[1], since it can handle with data having both numeric and categorical attributes.

- It has the same structure of the K-means algorithm.

- The novelty of this algorithm is the way it measures the distance between categorical data.

- When the data only is formed by numeric attributes, the algorithm reduces to the regular K-means algorithm

[1] Ahmad, A., Dey, L. (2007). A k-means clustering algorithm for mixed numeric and categorical data.Data & Knowledge Engineering 63. 503–527

Algorithm Description (2/4)

- Intuitively, we can see similarity in categorical data by finding common patterns among the different set of attributes so that these patterns characterise well the data in each cluster.

- These patterns are usually seen as how likely is the value of a given categorical attribute to co-occur with the values of the other categorical attributes.

- Distance function based on the overall distribution of any two values of any categorical attribute

Algorithm Description (3/4)

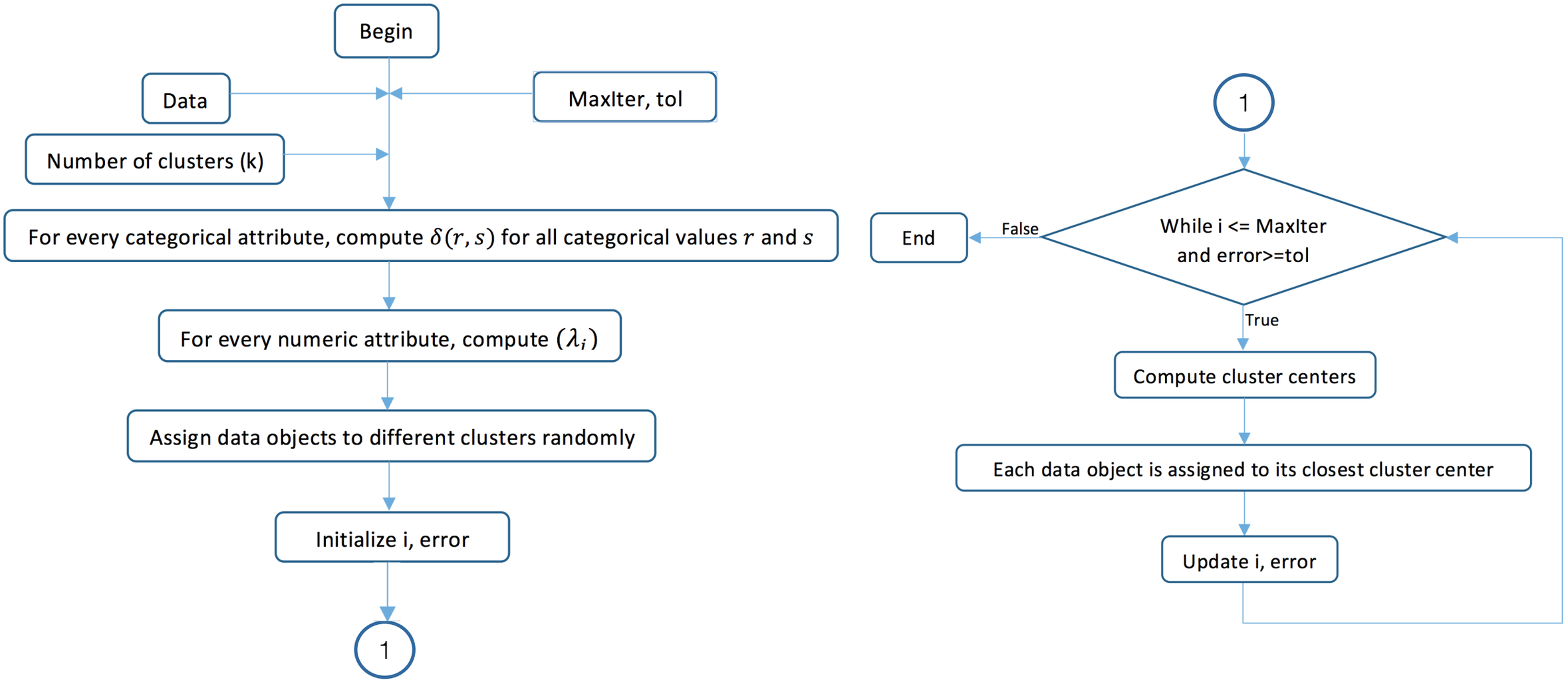
- Let ω be a subset of the support of the attribute A_j . Denote by $P_i(\omega|x)$ the conditional probability that an element having a value x for A_i has a value belonging to ω, then the distance of any two values x and y belonging to the attribute A_i respect the attribute A_j is defined as:

$$\delta^{ij}(x,y) = P_i(\omega|x) + P_i(\sim\omega|y) - 1$$

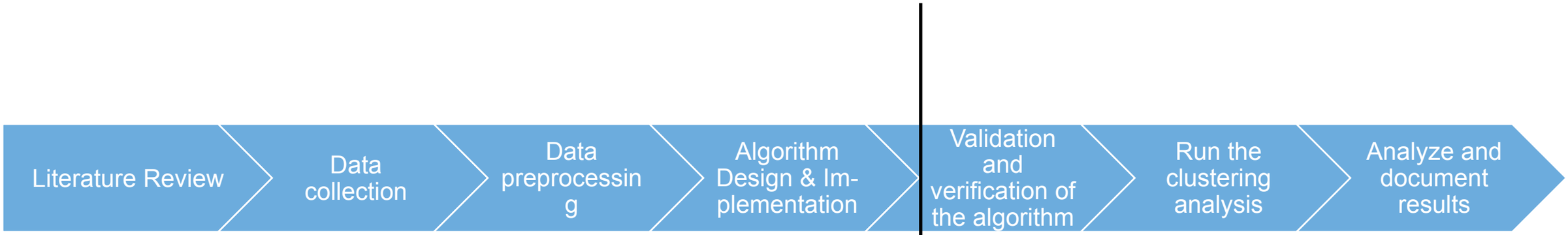
The distance of any two values x and y belonging to the attribute A_i is defined as the average of the distances respect the other attributes:

$$\delta(x,y) = \frac{1}{n_c - 1} \sum_{j=1,2,..,n_c\ j\neq i} \delta^{ij}(x,y)$$

Algorithm Description (4/4)



Summary


Literature Review
Data collection
Data preprocessing
Algorithm Design & Implementation
Validation and verification of the algorithm
Run the clustering analysis
Analyze and document results
Collected from the US census bureau
Includes:
- Cleaning
- Recoding
- Identifying the variables of interest
Having in mind:
- Data type
- Performance
- Theoretical properties

Thanks for your attention!

A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

Juan Sebastián Marín-Delgado
jmarind@eafit.edu.co
Tutor: Francisco Zuluaga-Díaz
fzuluag2@eafit.edu.co