A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

RESEARCH PRACTICE 3 PROPOSAL REPORT

JUAN SEBASTIÁN MARÍN DELGADO
TUTOR: FRANCISCO ZULUAGA

EAFIT UNIVERSITY
DEPARTMENT OF MATHEMATICAL SCIENCES
MATHEMATICAL ENGINEERING
MEDELLÍN
2015

PROBLEM STATEMENT

In response to the fast growing of the Hispanic population in the US in the last years, it has become imperative to develop a precision – targeting tool to market to Hispanics living in the US since they constitute the largest minority of this country. In fact, according to the projections done by the US Census Bureau, in 2050 one third of the US population will be Hispanics. Such tool should be designed to help understand the Hispanics characteristics and composition inasmuch as explaining Hispanic characteristics is a very complex task because Hispanic population is highly diverse, since Hispanic's cultural heritage comes from more than twenty nations and Hispanics have various levels of acculturation, ideals, literacy and affluence.

In order to understand the Hispanics household composition it becomes natural to think about identifying which are the most representative groups in the sense that the households in those groups have similar characteristics according to some variables of interest. In other words, we are interested in finding a set of groups characterized by the variables of interest so that for “almost” every household in the population it can be easily classified in one group in the set. Observe that the usefulness of the classification relies strongly on the fact that the groups are desired to be different enough, avoiding ambiguities and ensuring that if a household sample point can be classified, then there is one and only one group which contains that household sample point.

More formally, the problem described here can be presented as a clustering problem as follows:

Let $H = \{H_1, H_2, \dots, H_n\}$ be the set containing the Hispanic household sample of size n. Assume that each H_i in H is a p −dimensional vector containing the information of the p variables of interest, then the clustering of H is the partitioning of H into k clusters: $C_1, C_2, \dots, C_k$ so that they satisfy the following conditions:

$$\bigcup_{i=1}^{k} C_i = H \qquad (1)$$

$$\forall i = 1, \dots, k \;\; C_i \neq \emptyset \qquad (2)$$

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

Note that the latter conditions constitute the theoretical formulation for a clustering problem, however, when dealing with real data it might become non-pragmatic to require every data point to be classified in a given cluster because of the presence of possible outliers, and therefore, for our problem we relax the condition (1) to be $\bigcup_{i=1}^{k} C_i \subseteq H$. In addition, it should be noticed that our problem formulation is different from the classical theoretical formulation (that many authors present) since we do not assume $H_i \in \mathbb{R}^p$, if we did, one immediate consequence would be that all the attributes constituting each H_i were endowed with the notion of order existent in the set of real numbers, thus restricting the type of data to be used (i.e. nominal categorical data could not be used).

It is interesting to note that the second condition, aims to prevent the existence of empty clusters since they would not be useful and finally, the third condition guarantees that if a household sample point can be classified then its classification is unique (no more than one cluster contains that household sample point). In addition, one might want the set of clusters to be small enough so that they are practical and useful to understand the Hispanic household characteristics, but on the other hand, they should be adequately large to capture the diversity immersed in the Hispanic population.

It is also important to notice that for any given H there exist many different set of clusters which satisfy the last conditions, meaning that the solution for the clustering problem is not unique. Therefore, it becomes necessary to define a selection criteria which lead to get only one solution. This selection criteria usually depends on the researcher needs and the particular desired properties about the solution. However, one common approach consist of defining a fitness function which aims to measure the quality of the obtained cluster, an example of such type of functions that has been widely used is the total mean square error (MSE), which is defined as follows:

$$f(H,C)=\sum_{i=1}^{k}\sum_{H_j\in C_i}d^2(H_j,O_i) \qquad (4)$$

where O_i represents the centroid of the i-th cluster (C_i) and $d(.,.)$ is a distance function specifying how far is each data point from its cluster centroid. Thus under this approach we want to minimize f by minimizing the distance of each data point to its centroid. Other approaches use different kind of distance functions which instead of minimizing the distance of each data point to its centroid, aim to maximize the distance between clusters. Finally, it is also important to notice that the design of the fitness and distance functions depend strongly on the type of data. In our case, the household sample points will be formed by both numeric and categorical data thus the usual distance functions like the Euclidian distance, the infinity norm or the p-norm are not the most adequate since they cannot capture the notion of distance function for categorical data, nonetheless, they still can be useful for defining a new distance function which can measure adequately the distance between data points formed by a mix of numeric and categorical data.

GENERAL OBJECTIVE

Generate a useful classification of the Hispanic households in the U.S. to understand the Hispanic Household composition and its characteristics to support further marketing strategies.

SPECIFIC OBJECTIVES

1. Make a review of literature about the clustering algorithms
2. Prepare the Hispanic household data for the clustering analysis
3. Design and implement an adequate clustering algorithm for the Hispanic household data
4. Verify and validate the desired properties of the implemented algorithm (i.e. convergence)
5. Classify the Hispanic household data using the implemented algorithm and analyze the results.

REVIEW OF LITERATURE

There exist abundant literature exposing an array of different approaches to solve the clustering problem since it is an attractive and important task in data mining that is used in many applications. Due to this large variety of applications, different data types and various purposes it is difficult to find a unique algorithm that can fulfill all the requirements at once. According to (Tseng & Yang, 2001) clustering algorithms can be classified into two types: hierarchical and non-hierarchical. The hierarchical clustering algorithms recursively find clusters either in an agglomerative or a divisive way. The agglomerative ones merge together the most similar clusters at each level and the merged clusters will remain in the same cluster at higher levels. In the divisive methods, the initial stage view all the set of elements as a cluster and at each level, some clusters are binary divided into smaller clusters. On the other hand, the non-hierarchical methods find all clusters simultaneously without forming any hierarchical structures.

Although hierarchical methods have been used in different applications (including marketing and customer segmentation) as it can be seen in (B.Saglam, 2006; Bang & Lee, 2011; Hong, 2012; Hung & Tsai, 2008; Qin, Ma, Herawan, & Zain, 2014), non-hierarchical methods have shown to achieve better results, especially those which are center-based. One common example of this kind of methods is the K-means algorithm, which has become a remarkable algorithm for clustering problems because of its simplicity, easy implementation and its solutions quality (see Cheo, 2004; Jain, 2010; Kaufman & Rousseeuw, 1990)). It has been designed to minimize the intra-cluster variance (without ensuring that the result has a global minimum variance) (Kao, Zahara, & Kao, 2008; S.Z Selim & Ismail, 1984). Nonetheless, the K-means algorithm require to know in advance the number of clusters and is sensitive to the initial centroids (which can be given either by the user or chosen at random), making it likely to converge to local optima rather than global optima. Trying to overcome these issues, several heuristic methods have been developed; for instance, (S.Z. Selim & Alsultan, 1991) proposed a simulated annealing algorithm for the clustering problem, in (Arabia, 1995; Sung & Jin, 2000), it is presented a tabu search heuristic to conduct clustering. Genetic algorithms as well as Ant Colony Optimization heuristics have also been developed to perform clustering as can be seen in (Krishna & Murty, 1999; Maulik & Bandyopadhyay, 2000; Shelokar, Jayaraman, & Kulkarni, 2004; Tseng & Yang, 2001). Neural networks (self-organizing feature maps) have also been used to tackle the clustering problem, however, it is difficult to set up the training parameters and the computational time needed to run the algorithm is usually very high (Kuo, Ho, & Hu, 2002).

A more recent and novel approach has integrated the K-means algorithm with a powerful optimization heuristic called gravitational search algorithm which is inspired by the Newtonian Gravity Law (Hatamlou, Abdullah, & Nezamabadi-pour, 2012). This approach is particularly interesting since it takes the advantages of the K-means algorithm and makes it more robust and less sensitive to the initial centroids through the explore capabilities of the gravitational search algorithm, allowing the integrated algorithm to explore deeply the search space, thus making it more likely to converge to global optima rather than local optima. In (Hatamlou et al., 2012), several experiments were conducted to compare the quality of the results given by this algorithm with those achieved by other heuristics. The comparison showed that the integrated algorithm achieved better results in terms of the quality of the solutions and the convergence speed.

JUSTIFICATION

The project outcomes are important in the sense that they generate an impact mainly in two ways: first of all, the classification of the Hispanics household will provide technical arguments to support decision making and marketing strategies for Hispanics as well as will help understand the Hispanic household composition and characteristics; and second, the algorithms and methodology to be developed during the project will be useful for running future analysis to other minorities, or when new data be available and it becomes necessary to run the clustering analysis again.

SCOPE

It is important to notice that the clustering algorithms and their applications are their selves a whole research area. This work only focuses on implementing and applying an adequate clustering algorithm to classify the US Hispanic households and analyze the outcomes of the process. The pertinence of the work mainly relies on its usefulness not only for marketing purposes but also, in a more general fashion, to understand the Hispanics' characteristics.

PROPOSED METHODOLOGY

The review of literature was essentially important as it brought a big picture about what has been done in terms of clustering algorithms and its applications to customer segmentation, thus gave some insights on how to tackle the problem as well as key ideas for the design and implementation of the clustering algorithm. The next stage of the project will consist of all the data preprocessing and preparation, we will use the U.S Census data[1.] Note that this stage is particularly relevant in the project since it is not only concerned with the usual data preprocessing (like cleaning the data) but also identifying the variables that best characterizes and differentiates the Hispanic population from the rest of the population (which will be strongly related with the quality and usefulness of the final classification). In this stage it might also be needed an additional reduction of the dimensionality of the problem, if that is the case, proper statistical methods will be applied depending on the nature of variables.

The design and implementation of the algorithm would be conducted based on both the review of literature and the nature of the variables that will be taken into account for the clustering analysis. For instance, a heuristic method hybridized with the K-means algorithm seems to be a promising approach for the clustering algorithm development.

[1] U.S Census data is available at: http://www.census.gov/acs/www/data_documentation/about_pums

In order to test the algorithm and its properties, the quality of its solutions will be evaluated as well as its convergence characteristics. This stage might also include the design of different experiments (with simulated data) which help to verify the correct performance of the algorithm.

Once it is seen that the algorithm works properly, it will be used to run the clustering analysis for the Hispanic household data, the outcomes of the clustering process will be documented and analyzed.

Finally, it should be noted that for the data preprocessing and preparation, Access will be used; and the algorithm implementation would be conducted using R.

TIMELINE

Activity	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Week 14	Week 15	Week 16
Data preparation and preprocessing	▨	▨	▨											
Design and Implementation of the clustering algorithm			▨	▨	▨	▨	▨							
Verification and validation of the algorithm and its properties								▨	▨					
Execution of the clustering analysis for the US Hispanic households										▨	▨	▨		
Report write up and results discussion													▨	▨

INTELECTUAL PROPERTY

This project and its outcomes are property of Juan Sebastián Marín and Francisco Zuluaga in equal proportions.

BIBLIOGRAPHY

Arabia, S. (1995). A tabu search approach to the clustering problem, 28(9). Pattern Recognition. 29(3), 731–742. doi:10.1016/0031-3203(95)00022-R

B.Saglam. (2006). A mixed-integer programming approach to the clustering problem with an application in customer segmentation. European Journal of Operations Research, 173(3), 866–879.

Bang, Y.-K., & Lee, C.-H. (2011). Fuzzy time series prediction using hierarchical clustering algorithms. Expert Systems with Applications, 38(4), 4312–4325. doi:10.1016/j.eswa.2010.09.100

Cheo, C. (2004). Particle Swarm Optimization Algorithm and Its Application to Clustering Analysis, (1), 789–794.

Hatamlou, A., Abdullah, S., & Nezamabadi-pour, H. (2012). A combined approach for clustering based on K-means and gravitational search algorithms. Swarm and Evolutionary Computation, 6, 47–52. doi:10.1016/j.swevo.2012.02.003

Hong, C.-W. (2012). Using the Taguchi method for effective market segmentation. Expert Systems with Applications, 39(5), 5451–5459. doi:10.1016/j.eswa.2011.11.040

Hung, C., & Tsai, C.-F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Systems with Applications, 34(1), 780–787. doi:10.1016/j.eswa.2006.10.012

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. doi:10.1016/j.patrec.2009.09.011

Kao, Y.-T., Zahara, E., & Kao, I.-W. (2008). A hybridized approach to data clustering. Expert Systems with Applications, 34(3), 1754–1762. doi:10.1016/j.eswa.2007.01.028

Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons.

Krishna, K., & Murty, M. N. (1999). Genetic K-Means Algorithm. IEE Transactions on Systems, Man, and Cybernetics, 29(3), 433–439.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K -means algorithm for market segmentation, 29.

Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern Recognition, 33(9), 1455–1465. doi:10.1016/S0031-3203(99)00137-5

Qin, H., Ma, X., Herawan, T., & Zain, J. M. (2014). MGR: An information theory based hierarchical divisive clustering algorithm for categorical data. Knowledge-Based Systems, 67, 401–411. doi:10.1016/j.knosys.2014.03.013

Selim, S. ., & Ismail, M. A. (1984). K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(1), 81–87.

Selim, S. Z., & Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. Pattern Recognition, 24(10), 1003–1008.

Shelokar, P. ., Jayaraman, V. ., & Kulkarni, B. . (2004). An ant colony approach for clustering. Analytica Chimica Acta, 509(2), 187–195. doi:10.1016/j.aca.2003.12.032

Sung, C. S., & Jin, H. W. (2000). A tabu-search-based heuristic for clustering. Pattern Recognition, 33(5), 849–858. doi:10.1016/S0031-3203(99)00090-4

Tseng, L. Y., & Yang, S. B. (2001). A genetic approach to the automatic clustering problem, 34(October 1999).