A CLUSTERING APPROACH FOR US HISPANIC HOUSEHOLDS SEGMENTATION

RESEARCH PRACTICE 3 FINAL REPORT

JUAN SEBASTIÁN MARÍN DELGADO

TUTOR: FRANCISCO ZULUAGA

EAFIT UNIVERSITY

DEPARTMENT OF MATHEMATICAL SCIENCES

MATHEMATICAL ENGINEERING

MEDELLÍN

2015

TABLE OF CONTENTS

1. 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.

2.1 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.

2.2 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.

3. 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 (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). This method has been designed to minimize the intra-cluster variance (without ensuring that the result has a global minimum variance) (Kao, Zahara, & Kao, 2008; 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, (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 et al., 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.

4. 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.

5. 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 Hispanics' characteristics.

6. 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 consisted of all the data preprocessing and preparation, we used the U.S Census data[1]. Note that this stage was particularly relevant in the project since it was 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 was strongly related with the quality and usefulness of the final classification). In this stage was needed an additional reduction of the dimensionality of the problem, thus, principal components analysis and multiple correspondence analysis methods were applied depending on the nature of variables.

The design and implementation of the algorithm was conducted based on both the review of literature and the nature of the variables that were taken into account for the clustering analysis. The implemented clustering algorithm aimed to integrate both the categorical and numerical variables keeping the general structure of a K-means algorithm. In order to test the algorithm and its properties, the quality of its solutions

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

were evaluated as well as its convergence characteristics. This stage also included the design of different experiments (with simulated data), which helped to test the correct performance of the algorithm.

Once it was seen that the algorithm worked properly, it was used to run the clustering analysis for the Hispanic household data, the outcomes of the clustering process were documented and analyzed.

Finally, it should be noted that for the data preprocessing and preparation, SPSS (21) and Access (2010) were used; and the algorithm implementation was conducted using R Studio (0.98.1091).

7. ALGORITHM DESCRIPTION

The implemented algorithm was designed in order to generate solutions to the clustering problem, so that, given a data set and the desired number of clusters k, the algorithm partitions the data set into k clusters satisfying de conditions exposed in Section 2. The algorithm is based on the approach proposed by (Ahmad & Dey, 2007) since it can handle with data having both numeric and categorical attributes. Although there exist in the literature a few other approaches to handle mixed data for clustering purposes, the one exposed in (Ahmad & Dey, 2007) is particularly interesting since it tries to capture the notion of distance between categorical attributes in a novel and useful way. Observe that the Euclidean distance results very natural when dealing with numeric attributes not only because they are intrinsically endowed with a notion of order, but also because it defines a metric in the set of n-dimensional real numbers (satisfying non-negativity, the coincidence axiom, symmetry and the triangle inequality). On the other hand, there is no such natural or intrinsic notion of order and distance when working with some categorical data (e.g. nominal variables), even though they might have an order structure (e.g. ordinal variables) it is not enough to know how far is one category from another of the same attribute, thus the notion of distance does not become as natural as it is with numeric attributes.

In order to introduce and understand the approach used by (Ahmad & Dey, 2007) to measure the distance between two categorical values of the same attribute, let's first notice that in a good clustering solution we would expect the data in the clusters to be similar. Intuitively, we can see similarity in categorical data by finding common patterns among the different set of attributes so that these patterns characterizes 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 since this co-occurrence defines such pattern. Therefore, the distance measure between two different categorical values should take into account the co-occurrence of the different attribute's values, in other words, should be based on the overall distribution of any two values of any categorical attribute.

The latter approach not only is intuitive but also it can be easily proven that satisfies non-negativity, symmetry and reflexivity (respect to the class of distance 0), thus having very interesting theoretical properties which translates into good clustering outcomes. It is also important to note that the algorithm does not make any assumption about the distribution of the data and so the data need not to fit any particular known distribution.

Given the distance measures mentioned above for numeric and categorical data we can derive a new distance measure for mixed data straightforwardly by a linear combination of them.

More formally and according to (Ahmad & Dey, 2007), lets represent each data point as an n-tuple containing n_r numeric attributes and n_c categorical attributes such that $n_r + n_c = n$. Now consider a pair of arbitrary categorical attributes A_i, A_j $(i, j = 1,\ 2, \dots, n_c\ i \neq j)$ and let w be a subset of the support of the attribute A_j as well as $\sim w$ represent the complement of w with respect to the support of the attribute A_j. Denote by $P_i(w|t)$ the conditional probability that an element having value t for A_i has a value belonging to w, then the distance of any two values x and y belonging to the attribute A_i respect the attribute A_j is defined as:

$$\delta^{ij}(x, y) = P_i(\omega|x) + P_i(\sim\omega|y) - 1 \qquad (5)$$

where $\omega = max\,\{w \subset Supp(A_j)/P_i(w|x) + P_i(\sim w|y)\}$. Finally, 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 as seen in Equation (6):

$$\delta(x, y) = \frac{1}{n_c - 1} \sum_{j=1,2,\dots,n_c\ j \neq i} \delta^{ij}(x, y) \qquad (6)$$

From the definition given above, it can be easily shown that the distance function satisfies the following properties:

$$0 \leq \delta(x, y) \leq 1 \qquad (7)$$

$$\delta(x, y) = \delta(y, x) \qquad (8)$$

$$\delta(x, x) = 0 \qquad (9)$$

Note that from (7), $\delta(x, y)$ is bounded and satisfies non-negativity, from (8) satisfies symmetry and from (9) satisfies reflexivity under the subset of all pairs (x, y) such that $\delta(x, y) = 0$. With the previous definition we can now define the distance any pair of data points Γ, Δ (recall that each data point is represented as an n-tuple containing the first n_r numeric attributes and the last n_c categorical attributes) as follows:

$$d(\Gamma, \Delta) = \sum_{i=1}^{n_r} (\lambda_i\,(\Gamma_i - \Delta_i))^2 + \sum_{i=n_r+1}^{n} \delta^2(\Gamma_i, \Delta_i) \qquad (10)$$

where λ_i is a weighting factor for the $i\text{-}th$ numeric attribute, it is obtained through the discretization of the numeric attribute into S intervals. Let $u_i[r]$ the $r\text{-}th$ interval of the discretization of the $i\text{-}th$ numeric attribute, then λ_i is computed with the following formula:

$$\lambda_i = \frac{2}{S(S-1)} \sum_{k=1}^{S} \sum_{j>k}^{S} \delta(u_i[k], u_i[j]) \qquad (11)$$

The cluster centers are represented as an n-tuple, where the first n_r components are the averages of the numeric attributes and the last n_c components are also tuples containing as many components as categories each categorical attribute has. The components of the categorical tuples contain a natural number representing the number of sample points in a specific cluster containing the specified categorical value of the corresponding attribute. The algorithm follows the general structure of the k-means algorithm, as it can be seen in Figure 1.

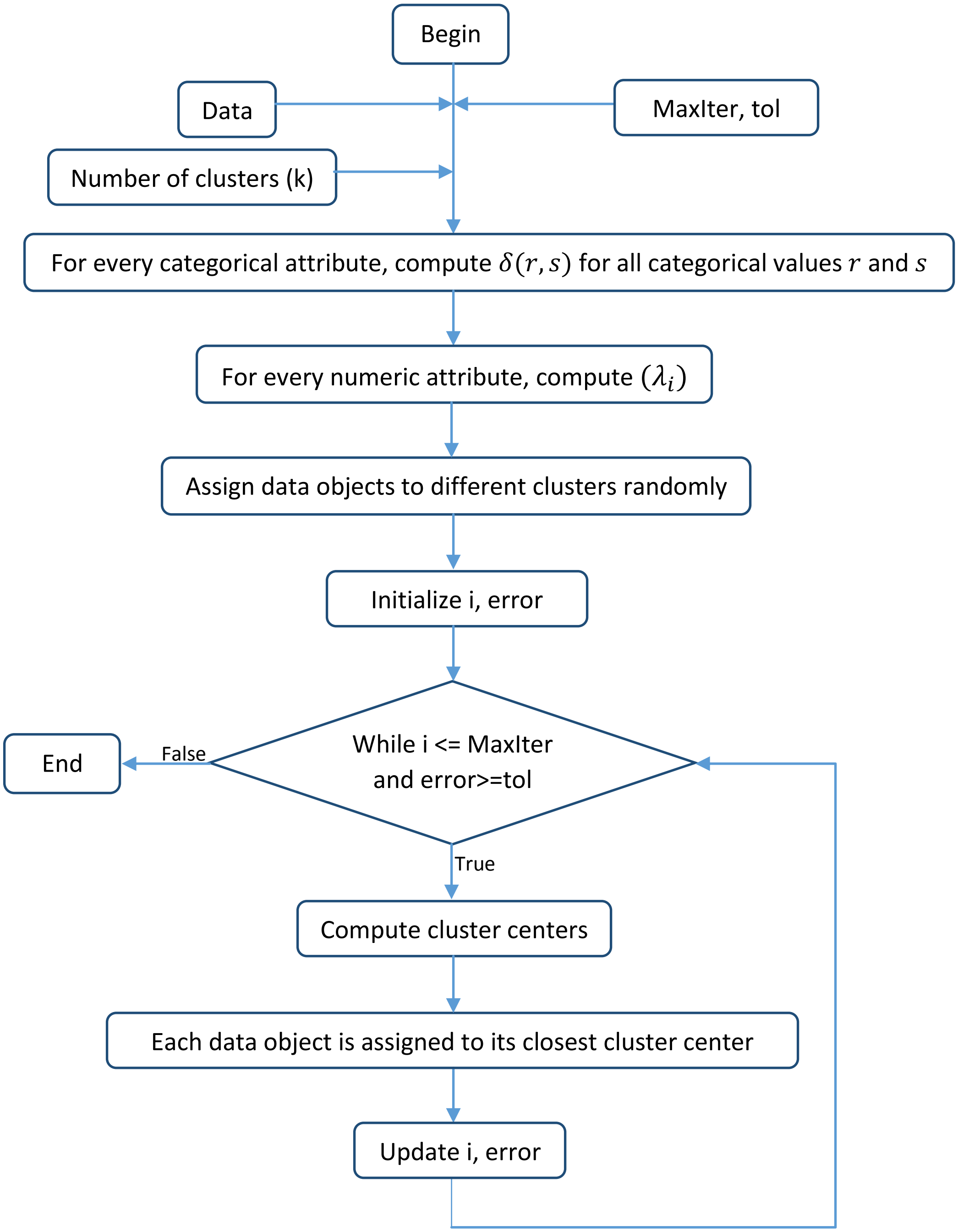

Begin
Data
MaxIter, tol
Number of clusters (k)
For every categorical attribute, compute $\delta(r, s)$ for all categorical values r and s
For every numeric attribute, compute (λ_i)
Assign data objects to different clusters randomly
Initialize i, error
While i <= MaxIter and error>=tol
False
End
True
Compute cluster centers
Each data object is assigned to its closest cluster center
Update i, error

Figure 1. Flow diagram of the implemented clustering algorithm

8. VALIDATION AND VERIFICATION

The algorithm was verified and validated using the same data sets that were used by (Ahmad & Dey, 2007) in their experiments. Those data sets have been widely used in the literature for evaluating the performance of clustering algorithms. Furthermore, the algorithm was tested with three of those data sets, each data set containing a different type of data, for instance: only categorical attributes, only numeric attributes and both numeric and categorical attributes. The evaluation method to measure the implemented algorithm performance is based on the proportion of data points belonging to the desired clusters.

Formally, let $T = \{T_1, T_{2,\dots,} T_k\}$ be the partition containing the natural (real) clusters of the input data and let $C = \{C_1, C_{2,\dots,} C_k\}$ be the output of the clustering algorithm for k clusters. Denote by $\#(\cdot)$ the cardinality of the set in its argument, naming p the measure of performance of the implemented algorithm we have that p can be calculated as follows:

$$p = \left(\sum_{i=1}^{k} \#(C_i \cap T_i)\right)\left(\sum_{i=1}^{k} \#(C_i)\right)^{-1} \qquad (12)$$

Tables 1, 2 and 3 show the results achieved by the implemented algorithm and the results reported by (Ahmad & Dey, 2007) for pure numeric data, pure categorical data and mixed data, respectively. The first data set is called Iris, containing 4 numerical attributes with 150 data points equally distributed into three different clusters. The second data set is named Vote formed by 16 categorical attributes and 435 elements split into two clusters (republicans and democrats), and finally the third data set has 690 data points with 14 variables out of which 8 variables are categorical and the other 6 are numeric.

Table 1. Iris data set comparative results

Algorithm	No. of data points in desired clusters	p
Implemented Algorithm	140	0.93
Algorithm proposed by (Ahmad & Dey, 2007)	142	0.95

Table 2. Vote data set comparative results

Algorithm	No. of data points in desired clusters	p
Implemented Algorithm	381	0.88
Algorithm proposed by (Ahmad & Dey, 2007)	377	0.87

Table 3. Australian Credit data set comparative results

Algorithm	No. of data points in desired clusters	p
Implemented Algorithm	591	0.86
Algorithm proposed by (Ahmad & Dey, 2007)	609	0.88

The tables above illustrate the average results of the implemented algorithm after 20 runs, and the results for the algorithm proposed by (Ahmad & Dey, 2007) after 100 runs. The statistical analysis let us conclude that there is no statistical difference in the p measure of both algorithms, which is clear evidence to validate the results and performance of the implemented algorithm. In addition, it is interesting to note that the performance achieved by the algorithms is above 0.85 in the examined cases for the three types of data showing its robustness when dealing with different kind of attributes.

9. CONVERGENCE OF THE IMPLEMENTED ALGORITHM

Convergence is critically important when developing any kind of algorithms or numerical methods since it guarantees that the algorithm will finish in a finite amount of time with the generated output satisfying the required numerical precision. Although we do not present a formal proof of convergence for the implemented clustering algorithm, we test its convergence through experimentation. Mathematically, the clustering algorithm is said to be convergent if there exist $n \in \mathbb{N}$ in the sequence of iterations $\{i_1, i_2, \ldots\}$, such that for every iteration $k \geq n$, $Sol(i_k) = Sol(i_n)$, where $Sol(i_j)$ represents the solution associated with the i_j-th iteration. Notice that the previous definition is just formalizing the fact that convergence of the algorithm is achieved once all the data points remain in the same clusters thus the solutions from a certain iteration and on are exactly the same.

Although it would be ideal to check convergence based on the definition given above, it is time consuming and adds significant computational complexity to the algorithm (both in time and memory), therefore we use an alternative way which is equivalent to the definition given above for the implemented algorithm. The alternative way consists of the sum of the intra-variances of each cluster, with the advantage that the process for computing it can be done in parallel with the clustering process hence not adding significant complexity. The sum of the intra-variances f for partitioning $\{x_i\}_{i=1}^{n}$ into k clusters $\{S_i\}_{i=1}^{k}$ is defined as follows:

$$f = \sum_{j=1}^{k} \sum_{x_i \in S_j} d^2(x_i, C_j) \qquad (13)$$

where d is as in (10). In this alternative approach the convergence is achieved once $\Delta f \equiv f_k - f_{k-1} = 0, k \geq 2$. However, since it can take long time Δf to be strictly zero, we relax that condition for practical purposes and redefine it to be $\Delta f < \mathrm{tol}$, where tol is a user predefined tolerance specifying the precision required by the user (it can also be understood as the maximum error allowed by the user). Figure 2 shows the behavior of the sum of intra-variances for the first 10 iterations for each data set used and described in the previous section.

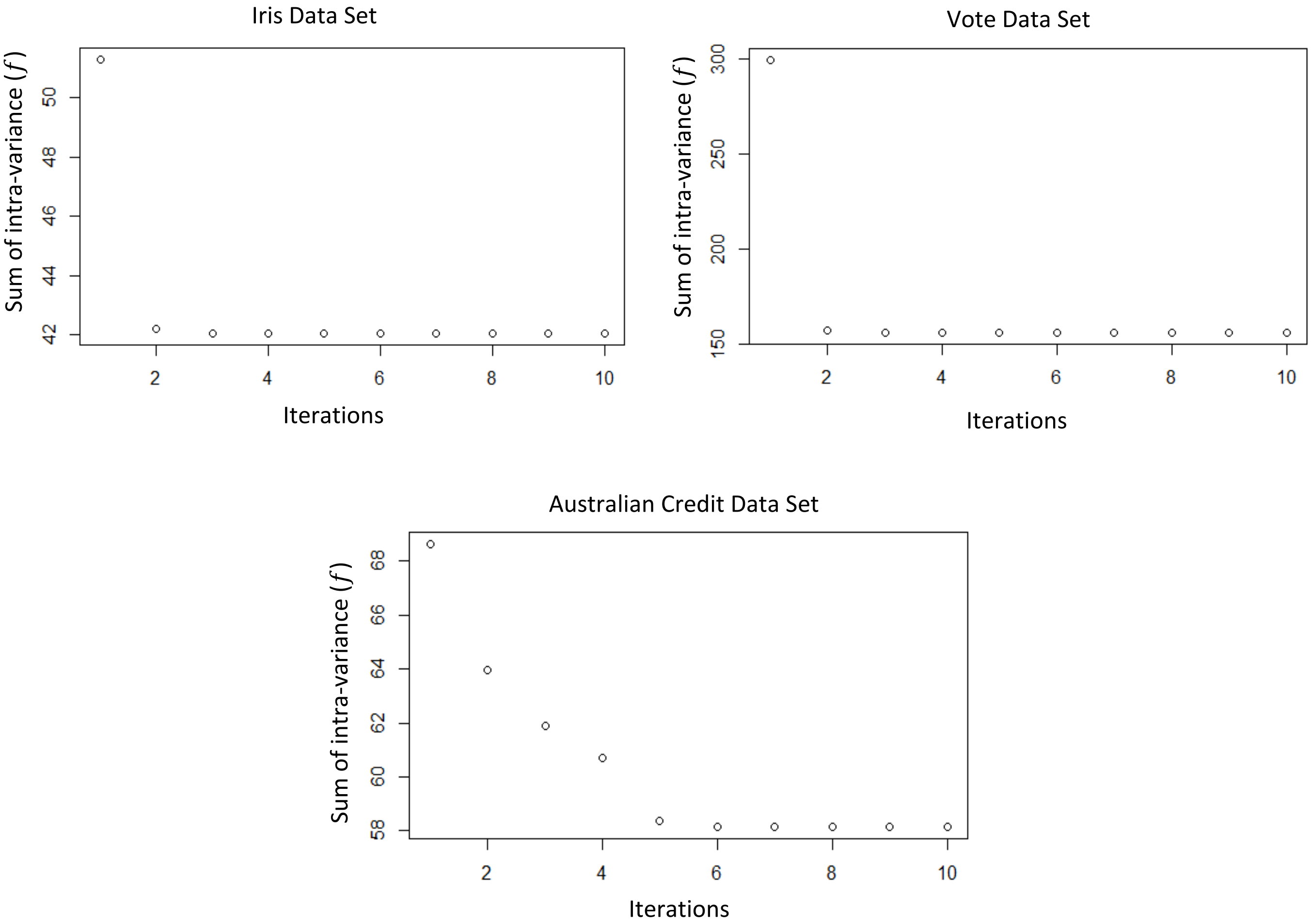


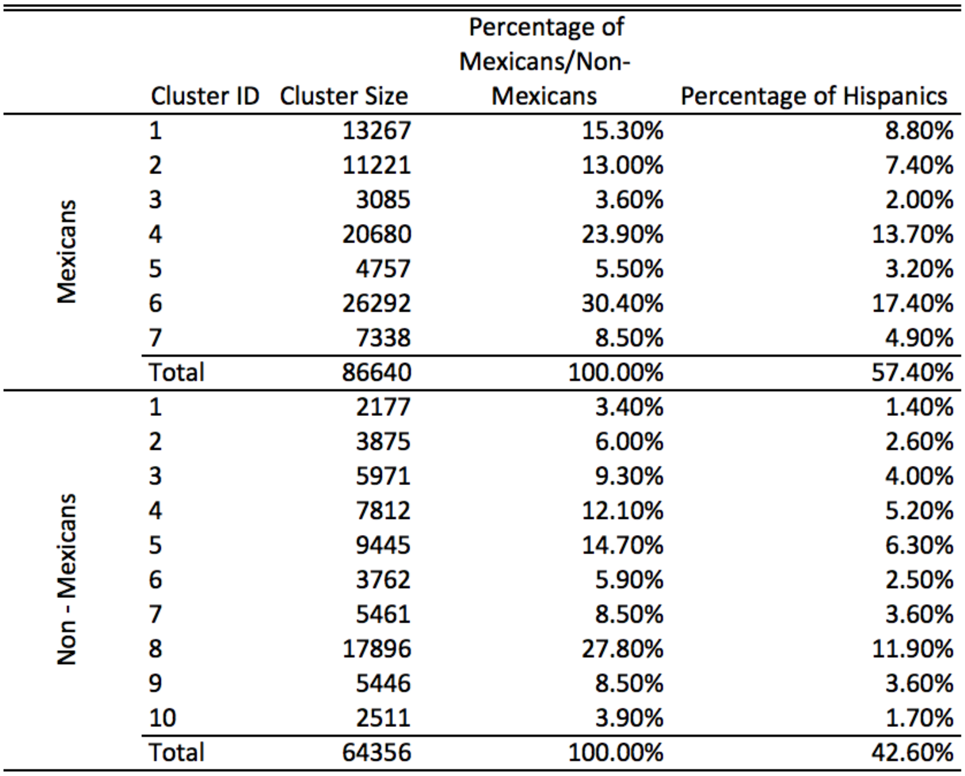
Figure 2. Convergence behavior of the implemented algorithm for each data set

From Figure 2 it can be seen the convergent behavior of the algorithm. It is also interesting to note that the sum of intra-variance is strictly decreasing function of the iterations thus converges monotonically. Notice the fast decreasing rate of the sum of intra-variance from one iteration to another, it is especially remarkable for the Iris and Vote data sets since they are formed only by one type of data. On the other hand, the decreasing rate for the Australian Credit data set still fast but not as fast as those in the other two data sets, this phenomena can be explained based on the fact that the Australian data has both numeric and categorical attributes, demanding more clustering effort which translates into a little bit slower convergence. For example, if we were to set the tolerance to be 0.2, then the algorithm would achieve the convergence in the third iteration for the Iris as well as the Vote data sets, whereas that would achieve the convergence in the sixth iteration for the Australian data set. Despite this natural difference in the convergence speed, the algorithm still converges quickly when dealing with mixed data.

10. RESULTS

In this section we present the results achieved by means of the application of the implemented clustering algorithm. In order to determine the optimal number of clusters, it was first conducted an exhaustive search varying the number of clusters from one to fifteen for both Mexicans and Non-Mexicans. The exhaustive search showed that the optimal number of clusters for Mexicans were seven and for Non-Mexicans were ten. Table 4 presents a general summary of the clusters density.

Table 4. Hispanics' clusters density



	Cluster ID	Cluster Size	Percentage of Mexicans/Non-Mexicans	Percentage of Hispanics
Mexicans	1	13267	15.30%	8.80%
	2	11221	13.00%	7.40%
	3	3085	3.60%	2.00%
	4	20680	23.90%	13.70%
	5	4757	5.50%	3.20%
	6	26292	30.40%	17.40%
	7	7338	8.50%	4.90%
	Total	86640	100.00%	57.40%
Non - Mexicans	1	2177	3.40%	1.40%
	2	3875	6.00%	2.60%
	3	5971	9.30%	4.00%
	4	7812	12.10%	5.20%
	5	9445	14.70%	6.30%
	6	3762	5.90%	2.50%
	7	5461	8.50%	3.60%
	8	17896	27.80%	11.90%
	9	5446	8.50%	3.60%
	10	2511	3.90%	1.70%
	Total	64356	100.00%	42.60%

Tables 5 and 6 show the main characteristics of the Mexicans and Non-Mexicans respectively. It was found that the characteristics that best differentiate the clusters were: the family type and employment status, the household language, whether or not the household was multigenerational, the presence and age of related children, the household ownership, the household income and lastly, the number of people living in the household. Notice that the first five variables are categorical while the household income and the number of people living in the household are numeric. The tables show both the variables and their respective categories (for the categorical ones) as well as the statistics used to characterize the clusters. It was also included in the table the geographic division to identify the most significant locations where the Hispanics are

Table 5. Non-Mexicans Clusters' Characteristics

Variable	Categories	Statistic	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Division	New England (Northeast region)	Proportion	5.2%	6.2%	8.1%	6.0%	4.5%	6.6%	7.7%	5.9%	7.6%	5.9%
	Middle Atlantic (Northeast region)	Proportion	24.3%	20.1%	22.8%	29.0%	18.1%	27.1%	22.4%	22.7%	21.0%	24.6%
	East North Central (Midwest region)	Proportion	4.6%	8.9%	5.7%	3.1%	7.1%	3.2%	5.7%	4.9%	5.8%	5.6%
	West North Central (Midwest region)	Proportion	0.7%	2.8%	1.1%	1.1%	2.5%	0.8%	1.8%	0.8%	1.7%	1.3%
	South Atlantic (South region)	Proportion	29.4%	19.9%	37.0%	21.0%	22.2%	20.5%	24.9%	39.4%	29.7%	24.5%
	East South Central (South region)	Proportion	0.6%	2.1%	1.4%	0.9%	2.3%	0.7%	1.9%	1.0%	1.7%	1.1%
	West South Central (South Region)	Proportion	7.3%	8.6%	8.4%	8.8%	8.1%	9.6%	10.0%	7.3%	9.5%	8.4%
	Mountain (West region)	Proportion	5.3%	11.7%	5.7%	3.1%	12.7%	3.6%	7.7%	7.9%	6.5%	6.5%
	Pacific (West region)	Proportion	22.5%	19.7%	9.9%	26.9%	22.5%	27.8%	17.9%	10.1%	16.4%	22.1%
Family Type and Employment Status	Married-couple family: Husband and wife in LF	Proportion	31.2%	42.5%	41.9%	21.0%	21.1%	39.3%	35.0%	17.6%	36.8%	28.2%
	Married-couple family: Husband in labor force, wife not in LF	Proportion	10.7%	13.9%	14.8%	8.5%	7.2%	16.2%	24.6%	7.6%	20.6%	13.8%
	Married-couple family: Husband not in LF, wife in LF	Proportion	5.3%	3.2%	3.2%	3.5%	4.4%	2.7%	2.0%	4.5%	1.7%	4.9%
	Married-couple family: Neither husband nor wife in LF	Proportion	5.6%	0.5%	1.3%	4.4%	8.5%	1.0%	1.3%	12.6%	0.4%	4.5%
	Other family: Male householder, no wife present, in LF	Proportion	5.1%	7.6%	6.2%	7.4%	2.4%	8.9%	7.3%	2.9%	11.5%	6.7%
	Other family: Male householder, no wife present, not in LF	Proportion	2.6%	0.9%	1.0%	1.1%	1.0%	0.9%	0.7%	1.4%	1.0%	2.0%
	Other family: Female householder, no husband present, in LF	Proportion	25.5%	25.4%	24.6%	10.1%	4.9%	26.7%	22.2%	5.5%	21.1%	28.8%
	Other family: Female householder, no husband present, not in LF	Proportion	14.0%	6.0%	6.8%	3.7%	2.9%	4.4%	6.9%	4.6%	6.9%	11.1%
	Not a family/NA	Proportion	0.0%	0.0%	0.0%	40.4%	47.7%	0.0%	0.0%	43.4%	0.0%	0.0%
Household Language	English only	Proportion	15.3%	90.8%	0.0%	6.6%	90.8%	3.2%	24.6%	0.0%	26.8%	20.7%
	Spanish	Proportion	81.5%	0.0%	99.9%	92.1%	0.0%	96.1%	72.5%	100.0%	69.8%	75.1%
	Other Indo-European languages	Proportion	1.9%	6.0%	0.1%	0.9%	5.7%	0.5%	1.9%	0.0%	2.1%	2.3%
	Asian and Pacific Island languages	Proportion	1.1%	2.5%	0.0%	0.3%	2.6%	0.2%	0.6%	0.0%	1.0%	1.6%
	Other language	Proportion	0.2%	0.7%	0.0%	0.2%	0.9%	0.0%	0.5%	0.0%	0.3%	0.3%
Multigenerational	No	Proportion	0.0%	100.0%	100.0%	99.8%	99.6%	100.0%	100.0%	99.9%	100.0%	0.0%
	Yes	Proportion	100.0%	0.0%	0.0%	0.2%	0.4%	0.0%	0.0%	0.1%	0.0%	100.0%
Presence and age of related children	Presence of related children under 6 years only	Proportion	0.0%	0.0%	0.0%	0.1%	0.2%	0.0%	0.0%	0.1%	100.0%	50.4%
	Presence of related children 6 to 17 years only	Proportion	80.4%	99.3%	100.0%	0.2%	0.4%	99.8%	0.0%	0.4%	0.0%	0.0%
	Presence of related children under 6 years and 6 to 17 years	Proportion	0.0%	0.7%	0.0%	0.0%	0.0%	0.1%	100.0%	0.0%	0.0%	49.6%
	No related children present	Proportion	19.6%	0.0%	0.0%	99.7%	99.4%	0.1%	0.0%	99.5%	0.0%	0.0%
Ownership	Owned with mortgage or loan (include home equity loans)	Proportion	55.9%	56.8%	47.0%	28.7%	39.9%	39.4%	38.1%	35.1%	34.6%	47.6%
	Owned free and clear	Proportion	13.1%	8.4%	8.4%	7.3%	19.1%	6.0%	5.1%	21.0%	4.3%	10.6%
	Rented	Proportion	30.4%	33.2%	43.5%	62.4%	39.3%	53.8%	55.2%	42.3%	59.5%	41.0%
	Occupied without payment of rent	Proportion	0.5%	1.6%	1.0%	1.5%	1.8%	0.8%	1.6%	1.6%	1.6%	0.8%
Income	-	Median	67,900	70,000	52,500	42,000	53,000	44,200	42,300	39,600	47,650	61,000
Number of P	-	Mean	4.09	2.23	3.27	2.14	1.25	3.77	4.03	1.73	2.77	4.62

Table 6. Mexicans Clusters' Characteristics

Variable	Categories	Statistic	C1	C2	C3	C4	C5	C6	C7
Division	New England (Northeast region)	Proportion	1.0%	0.3%	0.5%	0.4%	0.3%	0.3%	0.5%
	Middle Atlantic (Northeast region)	Proportion	2.5%	2.3%	1.0%	2.2%	1.6%	1.3%	2.4%
	East North Central (Midwest region)	Proportion	9.3%	8.3%	5.5%	8.1%	7.3%	6.2%	8.6%
	West North Central (Midwest region)	Proportion	4.1%	3.2%	2.0%	2.7%	2.5%	1.5%	3.8%
	South Atlantic (South region)	Proportion	6.1%	7.0%	3.2%	5.7%	4.8%	4.2%	7.8%
	East South Central (South region)	Proportion	1.8%	1.6%	1.1%	1.4%	1.0%	0.9%	1.7%
	West South Central (South Region)	Proportion	20.4%	26.9%	28.3%	27.8%	27.7%	36.4%	27.0%
	Mountain (West region)	Proportion	16.4%	13.1%	12.4%	12.9%	12.1%	12.9%	12.9%
	Pacific (West region)	Proportion	38.5%	37.3%	45.9%	38.9%	42.8%	36.2%	35.4%
Family Type and Employment Status	Married-couple family: Husband and wife in LF	Proportion	22.2%	31.3%	27.3%	40.0%	29.3%	17.8%	32.8%
	Married-couple family: Husband in labor force, wife not in LF	Proportion	7.2%	32.7%	15.9%	21.1%	19.8%	10.8%	24.5%
	Married-couple family: Husband not in LF, wife in LF	Proportion	4.3%	1.9%	5.3%	2.8%	4.4%	4.2%	1.7%
	Married-couple family: Neither husband nor wife in LF	Proportion	8.8%	1.3%	9.5%	1.4%	5.1%	11.9%	1.0%
	Other family: Male householder, no wife present, in LF	Proportion	2.7%	7.7%	5.2%	8.0%	6.5%	5.4%	13.0%
	Other family: Male householder, no wife present, not in LF	Proportion	1.2%	0.7%	2.9%	0.9%	2.1%	1.6%	0.9%
	Other family: Female householder, no husband present, in LF	Proportion	5.6%	17.2%	19.7%	20.9%	22.5%	6.3%	19.2%
	Other family: Female householder, no husband present, not in LF	Proportion	2.9%	7.3%	14.2%	4.9%	10.3%	4.4%	7.0%
	Not a family/NA	Proportion	45.1%	0.0%	0.0%	0.0%	0.0%	37.5%	0.0%
Household Language	English only	Proportion	95.3%	20.7%	16.4%	25.4%	18.6%	0.0%	31.6%
	Spanish	Proportion	0.0%	78.2%	81.2%	73.2%	79.0%	100.0%	66.5%
	Other Indo-European languages	Proportion	2.5%	0.5%	0.9%	0.7%	0.8%	0.0%	1.0%
	Asian and Pacific Island languages	Proportion	1.7%	0.4%	1.1%	0.5%	1.1%	0.0%	0.5%
	Other language	Proportion	0.5%	0.2%	0.4%	0.2%	0.5%	0.0%	0.3%
Multigenerational	No	Proportion	99.2%	100.0%	0.0%	100.0%	0.0%	99.7%	100.0%
	Yes	Proportion	0.8%	0.0%	100.0%	0.0%	100.0%	0.3%	0.0%
Presence and age of related children	Presence of related children under 6 years only	Proportion	0.1%	0.0%	0.0%	0.0%	41.0%	0.0%	100.0%
	Presence of related children 6 to 17 years only	Proportion	0.7%	0.0%	84.8%	100.0%	0.0%	0.1%	0.0%
	Presence of related children under 6 years and 6 to 17 years	Proportion	0.0%	99.9%	0.0%	0.0%	59.0%	0.0%	0.0%
	No related children present	Proportion	99.3%	0.0%	15.2%	0.0%	0.0%	99.9%	0.0%
Ownership	Owned with mortgage or loan (include home equity loans)	Proportion	40.2%	34.0%	51.2%	45.4%	46.2%	33.0%	30.6%
	Owned free and clear	Proportion	18.4%	10.1%	24.1%	12.8%	17.5%	26.8%	7.3%
	Rented	Proportion	39.2%	53.9%	23.5%	39.7%	34.8%	37.2%	59.4%
	Occupied without payment of rent	Proportion	2.2%	2.0%	1.3%	2.0%	1.4%	3.0%	2.7%
Income	-	Median	52,000	36,000	60,300	46,000	56,400	38,000	38,000
Number of People	-	Mean	1.32	4.66	4.49	3.6	5.39	1.99	3

settled, however it is important to note that even though it was not included as part of variables for the clustering analysis, it was mapped once the clusters were obtained.

The reason why the geographic division was not included in the clustering algorithm has to do with the fact that the Hispanic population is grouped around well known specific locations, so that adding this variable to the clustering algorithm would not enrich the clustering analysis, but certainly will increase the clustering effort.

From Tables 5 and 6 it can be seen that some determinant variables for defining the clusters are the household language, whether or not the household is multigenerational, the presence of related children, the household income and the number of people living in the household. Notice that the information contained in the tables summarizes the general characteristics of the found clusters and can be useful for identifying and targeting specific groups of Hispanics for any kind of purposes. In case that more variables need to be analyzed they can be easily mapped into each cluster and after doing so, determine if there is need or not of running another clustering analysis that include those variables. The characteristics presented in this work are useful in the sense that provide a general picture of the main groups of Hispanics and works well as a first step targeting tool which can be further refined to achieve more accurate results depending on the particular purpose of the user.

11. CONCLUDING REMARKS

In this work, we presented a clustering approach to tackle the problem of segmenting the US Hispanic households in order to find valuable information and patterns which are useful to characterize the subgroups of Hispanics living in the US and therefore the outcomes of the clustering procedure can be used as a targeting tool to get to specific groups of Hispanics depending on the needs and purpose of the user. We found 17 significant clusters, more specifically: 7 clusters for Mexican Households and 10 clusters for Non-Mexican Households. Based on the clustering outcomes we summarized the most important characteristics of the found clusters so that it is easy to get a good general idea of the description of the clusters and their particularities.

Finally, more refined and accurate clustering structures can be achieved depending on the particular purposes of the user; however, the results still are important and useful in the sense that provide practical insights to go deeper in any particular cluster as well as to understand better the Hispanics characteristics and how they are correlated. Future works can be oriented mainly on two ways: the first one consists in improving the computational efficiency of the implemented clustering algorithm by using parallel computing, the second one consists in exploring other distance measures for categorical that eventually lead to achieve better results.

12. INTELECTUAL PROPERTY AND CONFIDENTIALITY

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

BIBLIOGRAPHY

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

Arabia, S. (1995). A tabu search approach to the clustering problem, 28(9). Pattern Recognition. *28*(9), 1443-1451.

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. In *Networking, Sensing and Control, 2004 IEEE International Conference on* (Vol. 2, pp. 789-794). IEEE.

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. 1999 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. Procedings of IEEE, Vol.78, No.9, pp.1464-1480, 1990.

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. Pattern Recognition, 34(2), 415-424.