Fall 2020

Association Rules Patterns Discovery From Mixed Data

Welendawa Acharige Charith A. Elson

ASSOCIATION RULES PATTERNS DISCOVERY FROM MIXED DATA

by

WELENDAWA ACHARIGE CHARITH AKANANKA ELSON

(Under the Direction of Ionut Iacob)

ABSTRACT

Finding Association Rules has been a popular unsupervised learning technique for discovering interesting patterns in commercial data for well over two decades. The method seeks groups of data attributes and their values where their probability density of these attributes at the respective values is maximized. There are currently well-established methods for tackling this problem for data with categorical (discrete) attributes. However, for the cases of data with continuous variables, the techniques are largely focusing on categorizing continuous variables into intervals of interest and then relying on the categorical data methods to address the problem. We address the problem of finding association rules patterns in mixed data by using another unsupervised learning technique, clustering. The data attributes are organized into categorical and continuous attributes groups, and then we find the association rules patterns among attributes in each group that would satisfy the required probability density thresholds. We have implemented and tested our method, which produces very good results when used on real, mixed data.

INDEX WORDS: Frequent itemsets, Association rules, Clustering

2009 Mathematics Subject Classification: 15A15, 41A10, 68W99,62H30

ASSOCIATION RULES PATTERNS DISCOVERY FROM MIXED DATA

by

WELENDAWA ACHARIGE CHARITH AKANANKA ELSON

B.S., University of Colombo, Sri Lanka, 2011

M.A., University of North Carolina Greensboro, 2019

A Thesis Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ASSOCIATION RULES PATTERNS DISCOVERY FROM MIXED DATA

by

WELENDAWA ACHARIGE CHARITH AKANANKA ELSON

Major Professor:	Ionut Iacob
Committee:	Goran Lesaja
	Hua Wang

Electronic Version Approved:
October 2020

DEDICATION

This thesis is dedicated to my beautiful wife Ama Waidyarathna “The Love of my life” and my loving parents and brother.

ACKNOWLEDGMENTS

I wish to acknowledge Dr. Iacob, Dr. Lesaja and Dr. Wang for making this a possibility even after several years.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF SYMBOLS

- $\cap$ Intersection
- $\cup$ Union
- $\in$ An element of a set
- $\subset$ A subset of a set
- $|||$ Norm
- $||$ Cardinality

CHAPTER 1

INTRODUCTION

Association rules, a popular unsupervised learning method that has been around for almost two decades, is utilized in determining patterns in sales transactions in most sales-based databases. In these databases, you find the transactions with some values under different variables with high probabilities. For two sets of values, A and B, an association rule is defined as $A \Rightarrow B$. A is called the "antecedent" and B is called the "consequent" of the association rule. The probability calculations are done using two definitions named "support" and "confidence." The "support" for the given rule $A \Rightarrow B$, is the probability of the union of A or B, $P(A \cup B)$ and the "confidence" is the conditional probability of $P(A|B)$.

Clustering is another unsupervised learning method used to cluster data items with similar attributes. There are several popular methods for clustering such as K-means and hierarchical clustering. ClustOfVar, in Chavent et al, [5], is a package in R introduced for clustering variables instead of the data items.

In this study, we have proposed a novel method in which variable clustering is used to determine the association rules for both qualitative and quantitative variables. We started from the clustering with qualitative data, subsequently, we have extended the method for quantitative data as well. We have managed to establish an unbelievable relationship between distance measure in clusters and the confidence of the association rules, and the result was used to determine whether association rules have the desired confidence.

Since we could not identify an appropriate distance function for all types of variables, which relates to the confidence measure of the association rules, our method cannot find all possible association rules that include continuous variables since there could be a very large set of data to choose from. We believe addressing the above issue would be a good direction to further this study.

Our method will be a great contribution to the field of data mining, which is also known as the knowledge discovery of databases in which they look for useful patterns in databases to improve the efficiency and the effectiveness of the operation. Some of the fields that use data mining are banking, medicine, and entertainment such as Netflix.

CHAPTER 2

ASSOCIATION RULES

2.1 Association Rule Mining

Association rules have been a popular tool used in businesses. Association rules can provide valuable information such as 60% of people who buy comprehensive motor insurance also buy health insurance; 80% of those who buy music online, also buy books online. There is a diverse number of areas that employ association rules. A few of them are credit card transactions to study the transactions and predict what customer is likely to purchase; medical patient histories to detect increased risks of further complications using their past medical data.

Customer	Purchases
1	Tiling Cement; Tiles
2	Paint; White Spirit
3	Paint; Wallpaper; Plaster
4	Paint; Plaster; Tiling Cement; Tiles

Table 2.1: Purchase data for some customers

The basic objective of the association rules is to find the variables, $X = (X_1, X_2, X_3...., X_p)$, appearing mostly in the database. This is often used in a binary sense where the presence or absence of the variable is considered, that is $X_j \in \{0, 1\}$. This is known as "market basket analysis," since the observations are supermarket sales. x_{ij}, j^{th} item of the i^{th} transaction, is assigned 1 or 0 based on whether it was in the sale or not. Table 2.1 consists of items purchased by a customer at a supermarket. The original data in Table 2.1 can be transformed into binary format to do further analysis as showing in Table 2.2 .

The idea of association rules is to find a collection of values, $v_1, v_2, ...v_L$ for X, whose probability of occurring, $P(v_l)$ for $l = 1, 2....L$ is relatively high. But this probability will

Customer	Tiling Cement	Tiles	Paint	White Spirit	Wallpaper	Plaster
1	1	1	0	0	0	0
2	0	0	1	1	0	0
3	0	0	1	0	1	1
4	1	1	1	0	0	1

Table 2.2: Binary converted data of Table 2.1

nearly be too small for a reliable estimation when there are many variables and many values under each variable are present.

Therefore instead of finding single values with high probabilities, it makes more sense to find regions of X-space with high probabilities. Let S_j be the set of all possible values of the j^{th} variable (its support, which will be defined later), let $s_j \subset S_j$, a subset of all possible values. Now, finding association rules can be stated as finding subsets of variable values $s_1, ...s_p$ such that probability of each variable taking the values of its respective subset at the same time,

$$P\left[\bigcap_{i=1}^{p}(X_j \in s_j)\right] \tag{2.1}$$

is relatively higher, Friedman et al, [9]. The intersection of subsets $\bigcap_{i=1}^{p}(X_j \in s_j)$ is referred to as a conjunctive rule. For qualitative variables, the subsets are a list of nominal values and for quantitative variables, subsets are contiguous intervals. If the subset happens to be the whole set of values ($s_j = S_j$), X_j won't appear in the rule.

2.2 Market Basket Analysis

The probability calculated in (2.1) is not feasible for very large databases. Therefore to simplify (2.1) further, only two types of subsets are considered. Either s_j is a single

value of X_j, v_{0j}, or the entire set of values of S_j. Hence (2.1) is simplified as

$$P\left[\bigcap_{j\in J}(X_j = v_{0j})\right] \tag{2.2}$$

Let K be the number of all the values in all j variables. Therefore

$$K = \sum_{j=1}^{p}|S_j| \tag{2.3}$$

where $|S_j|$ is the number of distinct values in X_j. To indicate whether a certain value was present or not in a transaction, new K binary variables, $Z_1, Z_2....Z_K$ are introduced. This transforms (2.2) into finding a subset of integers $\kappa \subset 1, 2, ..., K$, making the probability,

$$P\left[\bigcap_{k\in\kappa}(Z_k = 1)\right] = P\left[\prod_{k\in\kappa}(Z_k = 1)\right] \tag{2.4}$$

high, Friedman et al, [9]. (2.4) gives the standard formulation of the market basket problem. From the items in $\bigcup_{i=1}^{p} S_j$, a set of items that κ refers to, is called an "item set." The number of dummy variables in the item set is called its "size." (This is supposed to be less than p). (2.4) can be estimated by taking the proportion of observations that satisfy (2.5).

$$\hat{P}\left[\prod_{k\in\kappa}(Z_k = 1)\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\kappa}(Z_k = 1) \tag{2.5}$$

Z_{ik} is the value of Z_k for the i^{th} case.

This is referred to as "Support" (a proper definition is provided later in the chapter), $T(\kappa)$ of the item set κ. An observation i is said to contain the item set K if z_{ik} is 1 for all k in κ.

In the mining of association rules, a bound for support is specified and all items sets K_l that can be formed from the dummy variables $Z_1, Z_2.....Z_k$ is sought.

$$\{K_l | T(K_l) > t\} \tag{2.6}$$

2.3 ALGORITHMS

The threshold t is adjusted so that (2.6) consists of only a small number of a fraction of all 2^K possible item sets. Agrawal et al, [2] introduce the "Apriori" algorithm to address (2.6). The algorithm computes the supports for all single item sets and removes the ones with support less than the threshold, t. The second pass takes the items of two with the ones that are already chosen in the previous pass and ones with the supports less than the threshold are removed. This process continues until the highest number of items with a support less than the threshold are selected.

Definition 2.1. *To speed up the process and the convergence, the set of items with higher support, κ, returned by the Apriori algorithm is selected into two disjoint subsets, A and B. Therefore, κ can be written as $\kappa = A \cup B$, and defined as an "association rule" and indicated as $A \Rightarrow B$. The items on left, A is called "antecedent" and the items on right, B is called "consequent."*

Definition 2.2. *The proportion of observations that are the union of A and B is defined to be the "support" of the association rule $A \Rightarrow B$. $T(A \Rightarrow B)$ is used for the support of $A \Rightarrow B$.*

This is as same as the support derived from the item set κ. In other words, it is the probability of observing $A \cup B$ in data items.

Definition 2.3. *The "confidence" of the association rule $A \Rightarrow B$, $C(A \Rightarrow B)$ is defined by the ratio between the support of $A \Rightarrow B$, $T(A \Rightarrow B)$, and the support of A, $T(A)$.*

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)} \tag{2.7}$$

This can also be interpreted as $P(B|A)$. "Expected Confidence" is defined as the support of the consequent $T(B)$, which can be estimated with $P(B)$.

For an example, consider the rule $\{Paint\} \Rightarrow \{White\ Spirit\}$ for item set $\kappa = \{Paint, Tiling\ Cement, Tiles, White\ Spirit, Wallpaper, Plaster\}$. Support is calculated to be .25, suggesting $\{Paint, White\ Spirit\}$ appears in 25% of the observations. The confidence is calculated to be .67, suggesting when $Paint$ was purchased, 67% of the time $White\ Spirit$ was purchased as well.

In association rules finding, the goal is to find the rules with supports and confidences with their corresponding thresholds, t, and c. Eventually, there will be association rules that satisfy the conditions $T(A \Rightarrow B) > t$ and $C(A \Rightarrow B) > c$.

So far, association rule mining was for qualitative variables, transformed into boolean dummy variables, that is, it only indicated whether an item was present or not in a transaction. Databases could have qualitative and quantitative attributes in other domains. In Srikant et al, [21] addressed the problem of quantitative association rule by fine partitioning the quantitative attribute and combining adjacent partitions as needed. In this, they introduced a measure of partial completeness to measure the information lost due to partitioning. This measure helps the user decide whether or not to partition. The possibility of this method generating too many similar rules is addressed by using an interesting measure "greater-than-expected-value" to identify the interesting rules in the output.

In Aumann et al, [3], an approach was introduced with algorithms involving two specific types, which are qualitative to quantitative and vice versa with a single attribute on the left-hand side. The right-hand side has the distribution properties of the quantitative variable. Yoda et al, [25] introduced some optimization-based approach in which a new measure named "gain" was introduced. Extensions to the method in Yoda et al, [25] were introduced in Brin et al, [4], but the rules were limited to one or two attributes. Mata et al, [18] proposed an algorithm to optimize the support of item sets on uninstantiated intervals on numeric attributes.

In Salleb-Aouissi, [19], QuantMiner, an algorithm capable of handling both qualitative

and quantitative algorithms while optimizing the quantitative attributes to mine association rule, was introduced. QuantMiner is focused on maximizing the gain of an association rule and it penalizes the variables with large intervals.

2.4 CURRENT WORK

In this work, we establish a surprising connection between the confidence of association rules and Jaccard similarity, a popular similarity/dissimilarity measure, to be used on the clustered variables (unlike the traditional observations clustering) to determine association rules of data with mixed variables. Firstly, we develop our method for binary and categorical variables and then extend it to mixed variables.

CHAPTER 3

CLUSTERING

Cluster analysis is the process of separating a collection of observations into groups called "clusters" in such a way that objects in the cluster are related to one another than the objects in different clusters. For this, values of the variables of the objects are used as characteristics and these characteristics are used to cluster the variables.

Cluster analysis is used to determine whether or not objects can be placed into subgroups that have substantially different properties. The second objective needs an assessment of the degree difference between the objects assigned to respective clusters.

There are two ways data can be grouped into clusters as hard and soft clustering.

Definition 3.1. *In hard clustering, each data point either belongs to a cluster completely or not.*

Definition 3.2. *In soft clustering, instead of putting a data point into a separate cluster, a data point is assigned to a distribution over all clusters. This way a data point has a fractional membership in several clusters.*

An important measure with cluster analysis is the degree of similarity between the objects in clusters. Any clustering method uses some definition of similarity to figure out what clusters the objects belong to. Similarities or dissimilarities (lack of similarity) can be represented in a form of $N * N$ matrix with (i, i') entry giving the similarity/dissimilarity between i^{th} and i'^{th} observations, where $i, i' \in \{1, 2, ...N\}$.

3.1 MEASURE THE DISTANCE BETWEEN TWO CLUSTERS

There are a few strategies to measure the distance between clusters.

- Single

  This method takes the shortest distance between an item in one cluster and an item

in the other cluster. This tends to produce elongated clusters or chains (similar items because of their similarity with intermediate items)

- Complete

  In this, the minimum longest distance between an item in one cluster and an item in the other cluster is considered. The method tends to join clusters with the approximately same diameter, producing compact clusters.

- Average

  This method uses the shortest average of distances between all pairs of items in two clusters. This is very sensitive to outliers and it tends to join clusters with small variances.

- Centroid

  The mean of all data items in each cluster is computed and called the centroid. Then the minimum distances between centroids are used. This is relatively easy to understand and to be used.

- Ward

  The minimum sum of squares of differences between the items in the two clusters is calculated for this. This puts together clusters with a roughly equal number of components and this method is sensitive to outliers.

3.2 K-MEANS CLUSTERING

K-means algorithm is one of the most popular clustering methods. It is used for all quantitative variables scenarios. Let x_{ij} be a data point from j^{th} variable (attribute) and i^{th} observation, where $i = 1, 2.....N$ and $j = 1, 2....p$. A dissimilarity measure used is the Euclidean distance between i^{th} and i'^{th} observations is given by

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2 \tag{3.1}$$

At the beginning of the algorithm, assign each observation to the cluster with the nearest centroid mean, which is with the least squared Euclidean distance. In the next step, recalculate the centroid means for a cluster of observations in each cluster. The process will continue until the assignments won't change.

K- means algorithm was used on the data in Table 3.1.

	A	B	C	D	E	F	G
Alex	5	0	0	0	2	1	2
Bruce	2	1	2	0	0	0	0
Chris	0	0	1	4	0	0	1
Don	0	0	2	0	1	1	2
Emil	6	2	2	0	1	1	2
Fred	4	0	0	2	0	0	2

Table 3.1: Sample dataset

The distance matrix for the sample data was calculated.

K was specified to be three and the cluster diagram in Figure 3.1 was generated.

3.3 HIERARCHICAL CLUSTERING

The initial configuration assignment and the number of clusters specified for the K-means clustering algorithm affects the results of the K-means algorithm, whereas hierarchical clustering only requires the user to specify a measure of dissimilarity between groups of observations, that is based on pairwise dissimilarities among the observations in the two groups. The clusters at a lower level merge themselves to create a cluster at a higher level, hence it has the name hierarchical clustering.

	Alex	Bruce	Chris	Don	Emil	Fred
Alex	0					
Bruce	.83333	0				
Chris	.83333	.8	0			
Don	.4	.83333	.6	0		
Emil	.33333	.5	.71428	.33333	0	
Fred	.6	.8	.5	.83333	.83333	0

Table 3.2: Sample distance matrix for the sample data

There are two main types of hierarchical clustering, Agglomerative (bottom-up) and Divisive (top-down). Agglomerative strategies begin from the bottom at every level some two clusters with the smallest dissimilarity are merged to form one cluster. As it gets to a higher level, the number of clusters reduces by one. The agglomerative method starts by considering all the observations as one big cluster and that splits into two clusters with the largest dissimilarity between those two groups. Eventually, there will be $N - 1$ level in the hierarchy.

A dendrogram is a graphical diagram that provides a highly interpret-able description of hierarchical clustering and this is one of the reasons hierarchical clustering is popular.

The data in Table 3.1 were clustered using hierarchical clustering. The dendrogram is shown in Figure 3.3.

3.4 Clustering of Variables

Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are two statistical tools used in multivariate data analysis for quantitative and qualitative variables respectively. As an alternative to PCA and MCA, the clustering of variables, even though cluster analysis was originally meant for clustering objects, can also be used to

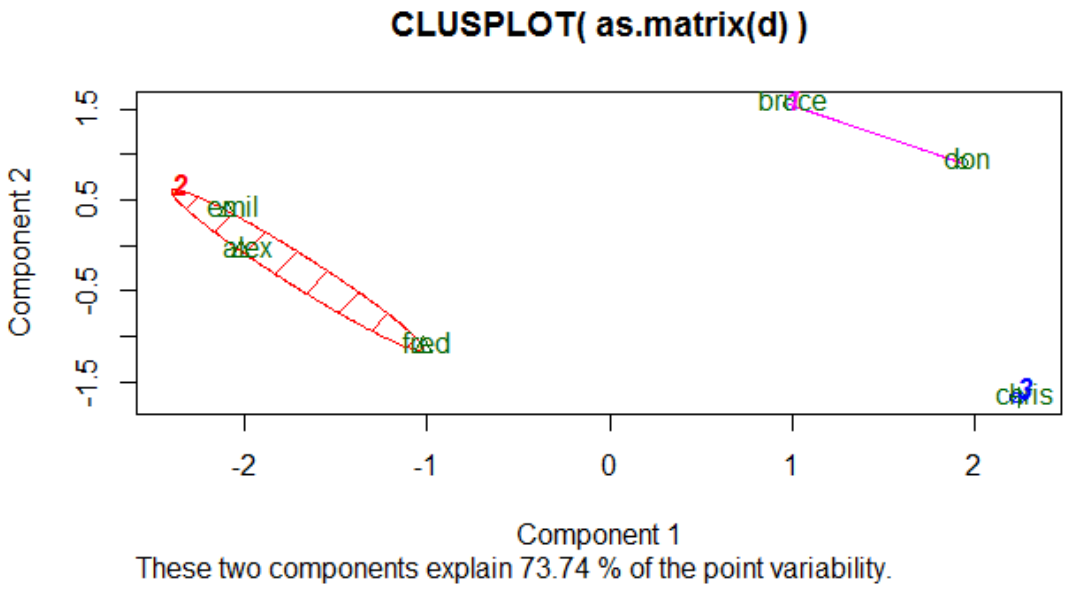

Figure 3.1: K-means clustering when K=3 for the data in Table 3.1

cluster variables into groups so that the meaningful structures can be derived. Variables that cluster together can be assumed to be strongly related to each other in a general perspective. Therefore when the variables are clustered together, selecting one variable from each group may be sufficient for carrying out the analysis. Perhaps one variable selected from each group could be synthetic for certain cases.

For clustering a set of variables, a common approach is to calculate the dissimilarity matrix between the variables and to apply a classical cluster analysis method used for clustering observations to the dissimilarity matrix. The functions in R to facilitate this are *hclust* from the package stats, introduced by Takeuchi et al, [22] and *agnes* from the package cluster, introduced by Maecheler et al, [17]. The type of dissimilarity matrix changes depending upon the fact whether variables are qualitative or quantitative. For qualitative variables, many measures can be used such as correlation coefficients (parametric or non-parametric) and as for qualitative variables measures such as Chi-squared, Rand, Belson, etc can be used. There are strategies to be used if the practitioner is not sure what measure

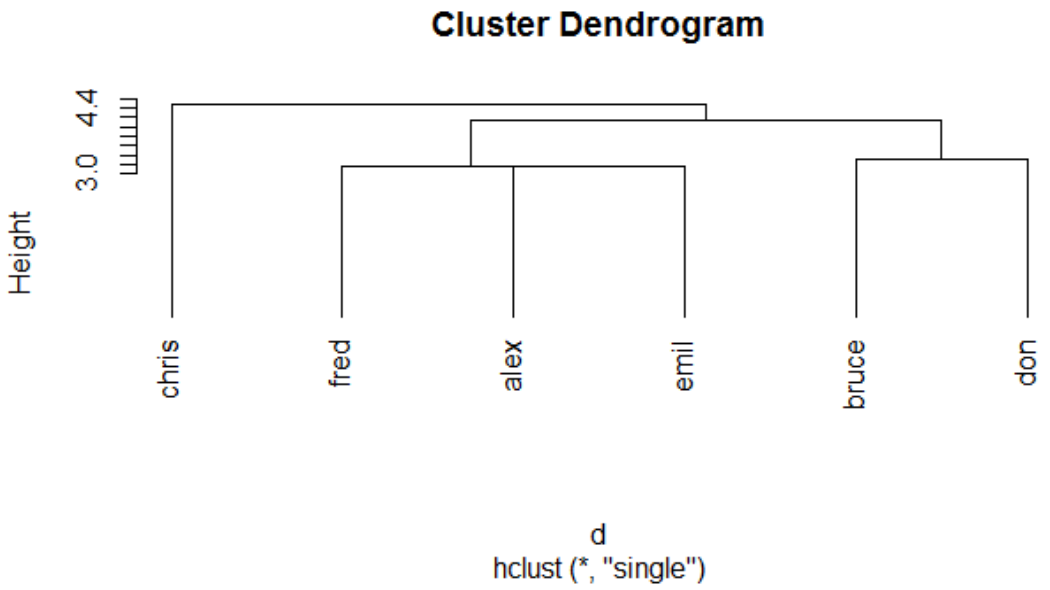

Figure 3.2: Hierarchical clustering for the data in 3.1

is to be used.

In the above methods, classical methods of clustering observations were transformed into the clustering of variables. There are also methods for directly clustering variables. VARCLUS in Sarle, W, S, [20] is such a tool developed by the SAS institute. Clustering around latent variables (CLV), a method based on PCA, and Diametrical clustering were introduced in Vigneau et al, [23], and Dhillon et al, [7] respectively. These methods are not implemented in other platforms than in R and they only work for quantitative variables.

A package for clustering of a mixture of both qualitative and quantitative variables has been introduced in R named ClustOfVar by Chavent et al, [5]. This also works excursively on qualitative or quantitative variables. In the package, two methods are proposed for clustering variables, a hierarchical clustering based algorithm and K-means based algorithm, which are used in functions *hclustvar* and *kmeansvar* respectively. These methods use PCAMIX, a PCA-based method for a mixture of qualitative and quantitative variables in Kiers et al, [13]. The ordinary PCA and MCA are two special cases that fall under

PCAMIX. A Singular Value Decomposition was performed on PCAMIX in Chavent et al, [6]. The clustering criterion they used is that variables were considered to be homogeneous when they have a strong relationship with a quantitative synthetic variable. The squared Pearson correlation and correlation ratio were used for quantitative and qualitative variables respectively.

Let $\{X_1, X_2....X_{p_1}\}$ be a set of p_1 quantitative variables and $\{Y_1, Y_2....Y_{p_2}\}$ a set of p_2 qualitative variables. Let X and Y be the corresponding quantitative and qualitative matrices of dimensions $n * p_1$ and $n * p_2$, where n is the number of observations. Let's denote $x_j \in R^n$ the j^{th} column of X and $y_j \in M^n$ the j^{th} column of Y with M the set of categories of y_j. Let $P_k = (C_1, ...C_K)$ be a partition into K clusters of the $p = p_1 + p_2$ variables.

The synthetic variable of a cluster C_k is $c_k \in R^n$ is defined to be the "most linked" to all the variables in C_k.

$$c_k = \arg\max_{u \in R^n} \Big\{ \sum\nolimits_{x_j \in C_k} r^2_{u,x_j} + \sum\nolimits_{y_j \in C_k} \eta_{u|y_j} \Big\} \tag{3.2}$$

where r^2 denotes the squared Pearson correlation and $\eta^2_{u|y_j} \in \big[0, 1\big]$ measures the part of the variance measured by the categories of y_j

$$\eta_{u|y_j} = \frac{\sum_{s \in M_j} n_s(\bar{u_s} - \bar{u})}{\sum_{i=1}^{n}(u_i - \bar{u}))^2} \tag{3.3}$$

where n_s is the frequency of category s, $\bar{u_s}$ is the mean values of u calculated on the observations belonging to category s and $\bar{u}$ is the mean of u.

The qualitative synthetic variable of a cluster is that when the first principal component when PCAMIX applied to all the variables in the cluster. These central synthetic variables are helpful in terms of reducing the dimension of the data. Further to clustering variables, the method is capable of evaluating the stability of the partition of variables and determining the number of clusters using the stability function.

CHAPTER 4

FINDING ASSOCIATION RULES OF MIXED TYPE DATA USING CLUSTERING OF VARIABLES

In this chapter, we present our association rule mining method of mixed data, based on the clustering of data variables and subsequently inferring interesting association rules-based on the clusters, we find and their proximities. Our main contribution is twofold: we perform data clustering on the dataset variables (rather than the traditional dataset samples), and we establish a relationship between the clustering distance and the association rules' confidence. This relationship allows us to find association rules from clusters of the dataset variables.

The chapter is organized as follows. We give some background on the association rules mining and clustering that are strictly specific for our work, as well as some notation in Section 4.1. In the rest of the section, we discuss the specifics of finding the association rules using clustering for binary data (Section 4.2), categorical data (Section 4.3), and continuous data (Section 4.4).

4.1 PRELIMINARIES

Let us start by establishing some notations. If $\mathcal{D}$ is a dataset of N samples of p dimensional data, we denote by x_i the sample $i = 1...N$ in $\mathcal{D}$, hence a vector of dimension $1 \times p$. We denote by $X_1, ..., X_p$ the variables in $\mathcal{D}$, which are also vectors of dimensions $N \times 1$. Consequently, a (scalar) data entry of $\mathcal{D}$ can be identified as both x_{ij} or x_{ji}. The dataset variables $X_1, ..., X_p$ can be binary (with two possible values, 0 and 1 or True and False), categorical (or qualitative, with a finite number of values), or continuous (or quantitative, when a continuous set of values is possible).

To establish the connection between data clustering and finding the association rules, let us consider a classic "shopping basket" dataset example, which is a classical application

#Items/Customers	A	B	C	D	E	F	G
Alex	5	0	0	0	2	1	2
Bruce	2	1	2	0	0	0	0
Chris	0	0	1	4	0	0	1
Don	0	0	2	0	1	1	2
Emil	6	2	2	0	1	1	2
Fred	4	0	0	2	0	0	2

Table 4.1: Shopping basket data example

of association rule mining. Table 4.1 shows such a dataset example, with six customers (Alex, Bruce, Chris, Don, Emil, and Fred) and seven products (A, B, C, D, E, F, and G). While the entries of each customer's shopping basket contain quantities, these quantities are irrelevant for the association rule mining. Rather, only the presence or absence of a product matters for the purpose of the association rule. However, for clustering one may consider these quantities as relevant to measure the similarities between different shopping baskets (depending on the distance measure being considered).

Let us next recall the definitions of the two main measures for association rules (ARs): support and confidence.

Definition 4.1 (AR support)**.** *Let $X \Rightarrow Y$ be an association rule, where X and Y are disjoint sets of data variables in a dataset $\mathcal{D}$. The* support *of the rule $X \Rightarrow Y$ is:*

$$sup(X \Rightarrow Y) = \frac{|X \cap Y|}{|\mathcal{D}|} \tag{4.1}$$

Definition 4.2 (AR confidence)**.** *Let $X \Rightarrow Y$ be an association rule, where X and Y are disjoint sets of data variables in a dataset $\mathcal{D}$. The* confidence *of the rule $X \Rightarrow Y$ is:*

$$conf(X \Rightarrow Y) = \frac{|X \cap Y|}{|X|} \tag{4.2}$$

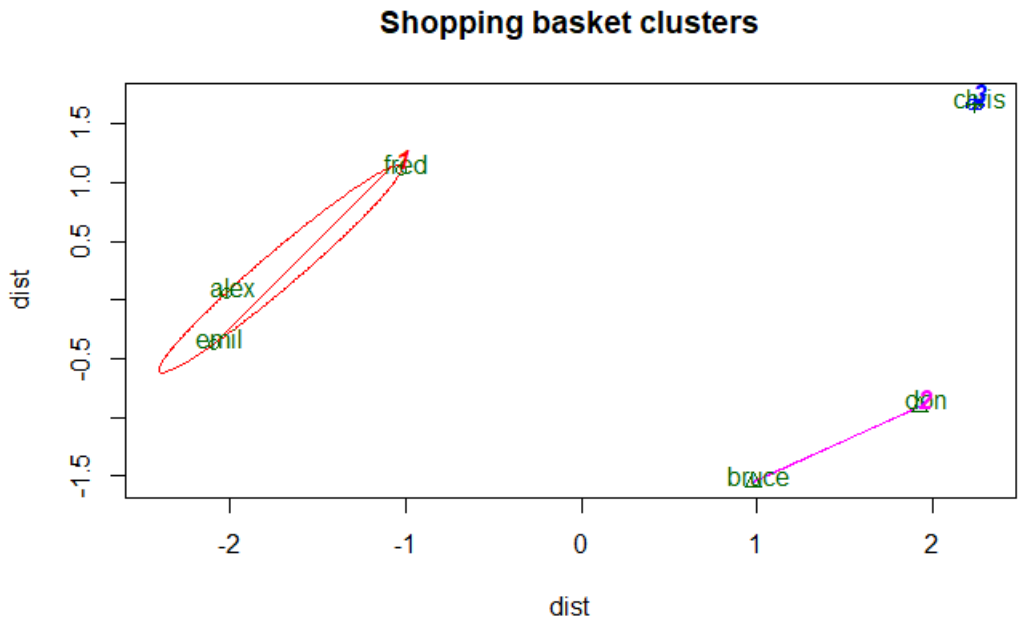

Figure 4.1: Clustering example for the shopping baskets data

For example, for the dataset in Table 4.1, the rule $\{A, B\} \Rightarrow C$ has the following support and confidence:

where the cardinalities union and intersection were computed as counts of the presence of one or the other, and presence of both products, respectively. An example of clustering of the dataset in Table 4.1 (using the Euclidian distance) is shown in Figure 4.1. We can immediately notice a discrepancy between finding the association rules and the clustering of the same data: while the former works with the data variables (columns), the latter computes similarities between the data customers (rows). To establish a connection between the two data analysis methods we need to address this discrepancy. Consequently, we will perform clustering of the data variables (columns), instead of the traditional data rows. Other significant differences between finding the association rules and clustering consist of the different measures (support, confidence, and distance) used for these analyses. We will need to find a connection between these different measures.

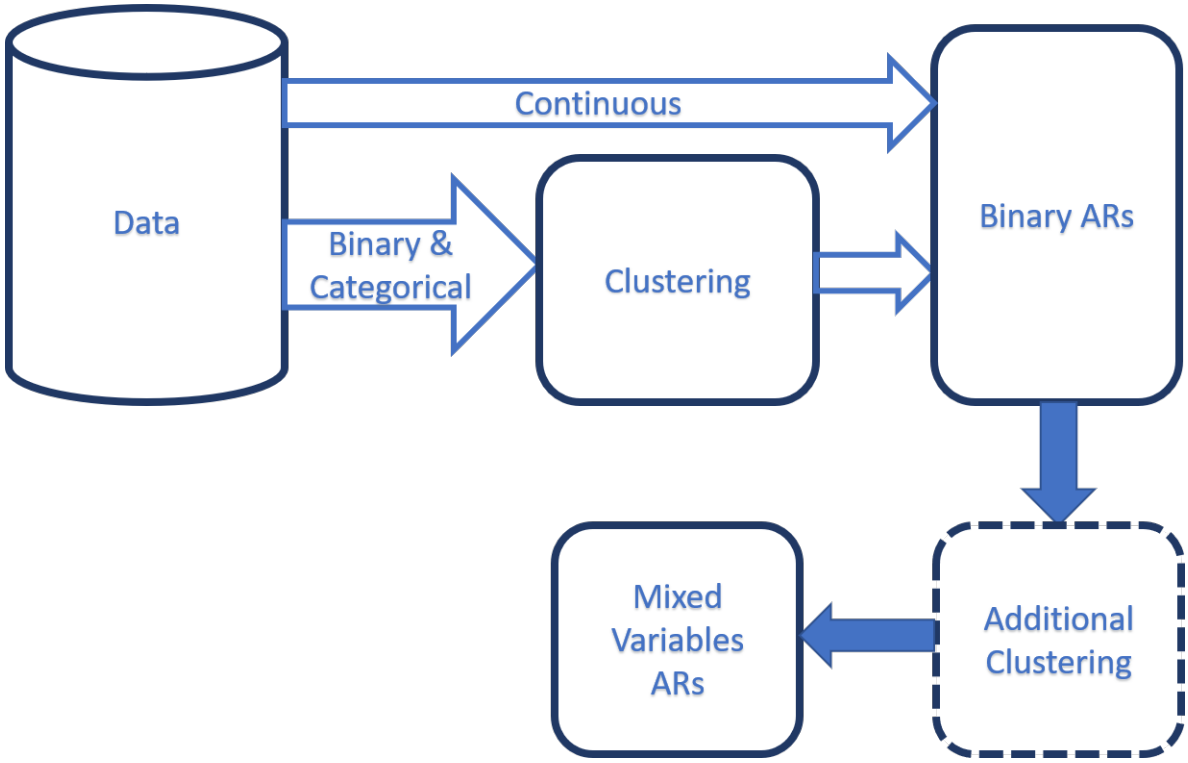

Figure 4.2: Mining Mixed Variables Association Rules using Clustering

The architecture we propose is illustrated in Figure 4.2. We first perform clustering on the binary and categorical variables using a similarity/distance measure as appropriate to determine the confidence of potential association rules (as described in the next section). Then we perform a linear search through the continuous variables set and include all continuous variables that satisfy the minimum required thresholds for support and confidence of the association rules. Additional clustering might be performed to find multiple continuous variables that can be included in the association rules.

4.2 The case of binary variables

We show that in the case of binary variables there is a close connection between a distance measure that can be used for clustering and the association rules' confidence.

The similarity measure defined below is widely used in many applications, such as

data mining and information retrieval (some references can be found in Kosub et al, [14]), or even similarities of DNA sequences, Vorontsov et al, [24].

Definition 4.3 (Jaccard similarity, Jaccard, P, [12])**.** *The Jaccard similarity coefficient of two sets A and B (not both empty) is defined as:*

$$J_{sim}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.3}$$

Clearly, the higher the coefficient value (between 0 and 1) the more similar the two sets are. For two binary variables X and Y, the Jaccard similarity between them can be quickly computed by using one norm and/or corresponding probabilities:

$$J_{sim}(X,Y) = \frac{||X \wedge Y||_1}{||X \vee Y||_1} = \frac{P(X \wedge Y)}{P(X \vee Y)}$$

For the corresponding Jaccard distance the following result is well-known:

Theorem 4.4 (Jaccard distance, Levandowsky el at, [15], Gilbert, G, [10], Lipkus, A, H, [16], Kosub, S, [14], Grygorian et al, [11])**.** *For the non-empty sets A and B, the function:*

$$J_{dist}(A,B) = 1 - J_{sim}(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{4.4}$$

represents a distance function.

The following result establishes a relationship between the Jaccard similarity/distance between two binary variables X and Y (or sets of binary variables) and the confidence of the corresponding association rule $X \Rightarrow Y$.

Theorem 4.5 (Relationship between Jaccard distance and confidence)**.** *Let X, Y be disjoint subsets of the dataset $\mathcal{D}$ set of variables, $X, Y \subset \{X_1, ..., X_p\}$, $X \cap Y = \emptyset$. Then:*

$$conf(X \Rightarrow Y) \geq J_{sim}(X,Y) = 1 - J_{dist}(X,Y) \tag{4.5}$$

Proof. The proof can be established immediately from the definitions of confidence and Jaccard similarity:

$$conf(X \Rightarrow Y) = P(Y \mid X) = \frac{P(X \wedge Y)}{P(X)}$$
$$= \frac{|X \cap Y|}{|X|} = \frac{|X \cap Y|}{|X \cup Y|} \cdot \frac{|X \cup Y|}{|X|}$$

Since:

$$\frac{|X \cup Y|}{|X|} \geq 1$$

it follows that:

$$conf(X \Rightarrow Y) = \frac{|X \cap Y|}{|X \cup Y|} \cdot \frac{|X \cup Y|}{|X|} = J_{sim}(X, Y) \cdot \frac{|X \cup Y|}{|X|} \geq J_{sim}(X, Y)$$

□

The theorem states that for two binary variables X and Y for which $J_{dist}(X, Y) < d$ we have that $1 - J_{dist}(X, Y) > 1 - d$ and hence $conf(X \Rightarrow Y) > 1 - d$. This guarantees that if the two variables X and Y are in a cluster with a diameter no larger than d then the association rule $X \Rightarrow Y$ has confidence larger than $1 - d$. Closer (more similar) two variables are, higher confidence is in their corresponding association rule.

For instance, for the variables A and C in the dataset in Table 4.1, we have:

$$J_{dist}(A, C) = \frac{|A \cap C|}{|A \cup C|} = \frac{2}{4} = .5$$

$$conf(A \Rightarrow C) = \frac{P(A \cap C)}{P(A)} = \frac{2}{4} = .5$$

$$conf(C \Rightarrow A) = \frac{P(C \cap A)}{P(C)} = \frac{2}{2} = 1$$

The result of Theorem 4.5 holds for the confidence of both rules $A \Rightarrow C$ and $C \Rightarrow A$.

The following result represents the fundament of our association rules mining using the clustering of binary variables (which is the first part of the diagram in Figure 4.2).

Theorem 4.6. *Let $X_1, ..., X_k$ be a cluster of k binary variables with diameter d: $max(J_{dist}(X_i, X_j) = d,\ i, j = 1, ..., k$. Then for any rule of the form $X \Rightarrow Y$ with $X, Y \subset \{X_1, ..., X_k\}$, $X \cap Y = \emptyset$ we have that:*

$$conf(X \Rightarrow Y) \geq 1 - d$$

Proof. The proof relies on the fact that for some $X_i \in X$, $Y_j \in Y$ we have that

$$J_{dist}(X_i, Y_i) \geq J_{dist}(X, Y)$$

hence

$$1 - J_{dist}(X_i, Y_i) \leq 1 - J_{dist}(X, Y)$$

Then by Theorem 4.5:

$$conf(X \Rightarrow Y) \geq 1 - J_{dist}(X, Y) \geq 1 - J_{dist}(X_i, Y_j) \geq 1 - d$$

Since $J_{dist}(X_i, Y_j) \leq d$, it follows that $1 - J_{dist}(X_i, Y_j) \geq 1 - d$ and therefore:

$$conf(X \Rightarrow Y) \geq 1 - d$$

□

An immediate consequence of Theorem 4.6 is that any association rule with variables from clusters of diameter at most $1 - \alpha$ will have a confidence of at least α. We, therefore, perform clustering of variables for a given dataset (using, for instance, compact hierarchical clustering), cut the hierarchy at a given distance $1 - \alpha$ (for a given parameter α), and all association rules from the resulting clusters have at least confidence α.

Two key observations are worth noting:

1. Some association rules may not have the desired support. The support threshold for each rule must be separately verified.

2. Some rules may not be discovered. Because (4.5) represents an inequality, rules with given confidence may be discovered if the diameter of a cluster less than $1 - \alpha$ is being considered. If this is important, in practice one can start at a fraction of the distance $1 - \alpha$ and subsequently verify if all confidences will pass the threshold α.

In the subsequent sections, we will explain how we deal with the cases of categorical and continuous variables.

4.3 THE CASE OF CATEGORICAL VARIABLES

Categorical variables can be considered an extension of binary variables by expanding each categorical variable into several binary variables, one binary variable for each category value. For instance, let us assume that an 8th categorical variable $\vec{H} = (a, b, a, a, c, d)$ is appended to the shopping basket dataset in Table 4.1 (as a column). We assume that the new categorical variable can take one of the discrete values $\{a, b, c, d\}$. Then H can be expanded in 4 binary variables as follows:

$$\vec{H}_a = (1, 0, 1, 1, 0, 0)$$

$$\vec{H}_b = (0, 1, 0, 0, 0, 0)$$

$$\vec{H}_c = (0, 0, 0, 0, 1, 0)$$

$$\vec{H}_d = (0, 0, 0, 0, 0, 1)$$

where a 1 or a 0 denotes the presence or absence of the respective category at the respective position. This approach is intuitive and easy to implement, however, an explicit expansion can considerably enlarge a dataset (especially if there are many categories) and consequently make computations significantly more expensive. In practice, however, the expansion needs not to be performed explicitly. The new binary variables $\vec{H}_a$, $\vec{H}_b$, $\vec{H}_c$, and $\vec{H}_d$ can be considered for computing desired cardinalities of unions and intersections

based on the values in $\vec{H}$, without explicitly creating their content. For instance, computing $||A \vee \vec{H}_a||$ amounts to counting rows where A is non-zero and $\vec{H}$ holds value a.

4.4 THE CASE OF CONTINUOUS VARIABLES

The case of continuous variables is difficult to address in practice and many solutions have been proposed in the literature (discretization and model as categorical/binary variables, heuristic methods, etc.). Typically, choosing one approach or another greatly depends on the data being analyzed and the practical application being considered. There is no "measure fits all" solution and no "best method" among the proposed solutions. In this section, we describe our approach, which adds to the multitude of the proposed solutions.

Let us consider the set of continuous variables $\{\vec{W}_1, ..., \vec{W}_p\}$. Given a desired confidence α, our method consists of the following three steps.

1. Find all binary variables clusters of diameter at most $1 - \alpha$.
2. For each continuous variable $\vec{W}_i = (w_{i1}, ..., w_{iN})$ and binary cluster $X_1, ..., X_k$
   - compute a range $[minW_i, maxW_i]$, where $minW_i$ and $maxW_i$
   - create a corresponding binary vector $\vec{b}W_i = (bw_{i1}, ..., bw_{iN})$, where

$$bw_{ij} = \begin{cases} 1 & \text{if } w_{ij} \in [minW_i, maxW_i] \\ 0 & \text{otherwise} \end{cases}$$

3. For each binary cluster, append the corresponding binary variables obtained from the continuous variables, then perform clustering again. Note that each continuous variable $\vec{W}_i$ must appear in each rule together with the range $[minW_i, maxW_i]$ found in the previous step.

For instance, let us consider an additional continuous variable for the dataset in Table 4.1: $\vec{W} = (1.2, -2.3, 1.1, 2.2, 2.0, 2.3)$. For the cluster $\{\vec{A}, \vec{C}\}$ with diameter .5 we considered

in Section 4.2, we compute $\vec{A} \wedge \vec{C} = (0, 1, 0, 0, 1, 0)$ and subsequently compute a $minW = -2.3$ and $maxW = 2.0$ (only among the second and fifth positions in $\vec{W}$). We then create the binary vector $\vec{b}W = (1, 1, 1, 0, 1, 0)$, with zeros for entries outside the range $[-2.3, 2.0]$ and ones otherwise. Next, we cluster (find the association rules) among the variables $\vec{A}, \vec{C}$, and $\vec{b}W$.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 THE CENSUS DATA

We have used the USCensus1990 dataset [8] for experiments and testing our method. The USCensus1990 dataset is a discretized version of the USCensus1990raw dataset. Many of the less useful attributes in the original dataset have been dropped, the few continuous variables have been discretized and the few discrete variables that have a large number of possible values have been collapsed to have fewer possible values.[1]

The USCensus1990raw dataset was obtained from the U.S. Department of Commerce Census Bureau website using the Data Extraction System. This system can be found at http://www.census.gov/DES/www/des.html.

The USCensus1990raw dataset contains a one percent sample of the Public Use Microdata Samples (PUMS) person records drawn from the full 1990 census sample (all fifty states and the District of Columbia but not including "PUMA Cross State Lines One Percent Persons Records"). A description of the fields and the coding of the values can be found in the Appendix. Additional information can be found at the Census Bureau website described above.

5.2 VARIABLE CLUSTERING AND FINDING ASSOCIATION RULES FOR THE CENSUS DATA

For all the experimental results we present here we used a PC equipped with an Intel Core i7-4770 CPU @3.40GH. The R code listing for producing our results are included in the Appendix.

We have randomly selected 20% of the Census data rows (about half a million rows)

[1] Unlike the USCensus1990raw dataset, the order of the cases in the USCensus HAS been randomized.

for our experiments. The dendrogram of complete hierarchical clustering is displayed in Figure 4.2.

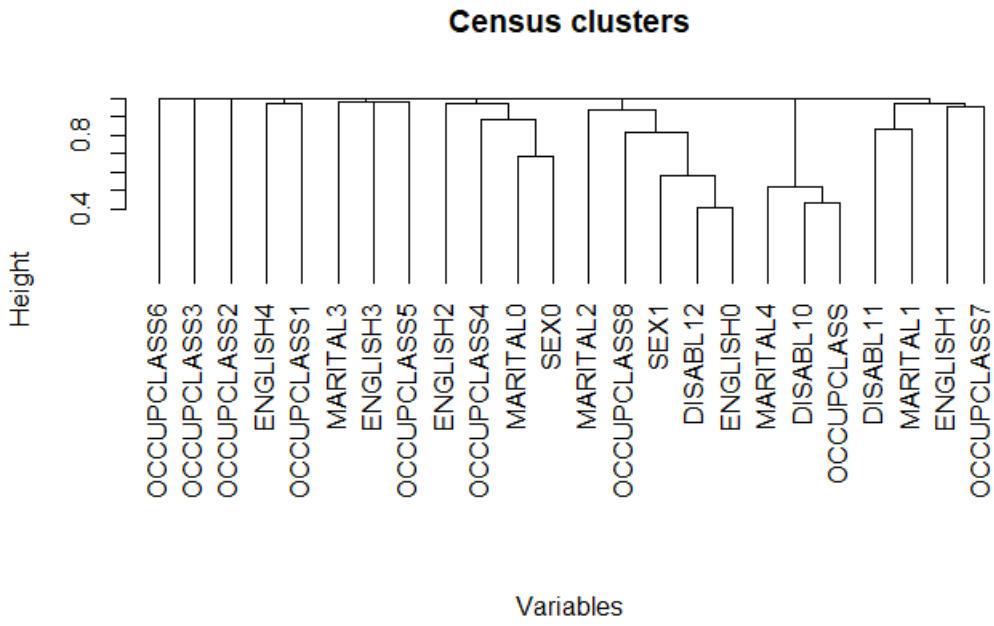

Figure 5.1: Clustering for a fragment of Census data (binary variables)

Association Rule	Cluster	Confidence
{DISABL10, MARITAL4} ⇒ {OCCUPCLASS}	1	0.5615501
{DISABL10} ⇒ {OCCUPCLASS,MARITAL4}	1	0.9995998
{DISABL12} ⇒ {ENGLISH0,SEX1}	3	0.4515397
{ENGLISH0, SEX1} ⇒ {DISABL12}	3	0.3201189
{DISABL10, DISABL12} ⇒ {OCCUPCLASS,ENGLISH0,SEX1}	1 and 3	0

Table 5.1: ARs confidences for Census data

We used a hierarchy cut at distance 0.7 and the resulting list of clusters of variables is given in Table 5.2. From the dendrogram, in Figure 5.1 we can determine that clusters 1 and 3 (for instance) are both below the cutting line, hence all pairwise distances are not

larger than $d = 0.7$. It follows that the result of Theorem 4.6 will apply to all association rules constructed with combinations of variables within each cluster, or between the two clusters. That is each such rule $X \Rightarrow Y$ must satisfy:

$$conf(X \Rightarrow Y) \geq 1 - d = 0.3$$

The results of the experiments are summarized in Table 5.1. All rules consisting of variables from the same cluster satisfy the result of Theorem 4.6. The last association rule in the table is composed of variables from different clusters, 1 and 3, which are joined into their super cluster above the height at $d = 0.7$. Hence it comes as no surprise that the confidence of such a rule is not greater than $1 - d = 0.3$, as it does not satisfy the requirements of the theorem.

	Variable	Cluster
1	DISABL10	1
2	MARITAL4	1
3	OCCUPCLASS	1
4	DISABL11	2
5	DISABL12	3
6	ENGLISH0	3
7	SEX1	3
8	ENGLISH1	4
9	ENGLISH2	5
10	ENGLISH3	6
11	ENGLISH4	7
12	MARITAL0	8
13	SEX0	8
14	MARITAL1	9
15	MARITAL2	10
16	MARITAL3	11
17	OCCUPCLASS1	12
18	OCCUPCLASS2	13
19	OCCUPCLASS3	14
20	OCCUPCLASS4	15
21	OCCUPCLASS5	16
22	OCCUPCLASS6	17
23	OCCUPCLASS7	18
24	OCCUPCLASS8	19

Table 5.2: Clusters of variables from the Census data

CHAPTER 6

CONCLUSION

Mining the association rules was initially introduced in the early nineties by Agrawal et al, [1], and it has been intensively studied ever since. Originally introduced for binary data, association rules mining for continuous data was capturing the attention of the data science research community shortly thereafter. However, while many approaches and algorithms were proposed, there is no measure fits all for finding the association rules for continuous data. It comes as no surprise that the subject is still of interest nowadays.

In this work, we proposed a novel method for the association rules mining for mixed data case: data that contains binary, categorical, and continuous variables. Our method relies on performing clustering of data variables, then inferring the association rules based on the clusters we find and the proximities between these clusters.

Our main contributions can be summarized as follows:

- We establish a surprising connection between one of the association rules' main analysis measure (confidence) and a popular similarity/distance measure (Jaccard similarity/distance) used for discrete (categorical) data.
- We perform clustering of data variables (columns), instead of the traditional approach where rows are used.
- We introduce a novel method of finding association rules based on the clustering of data variables. First, we develop our method for binary and categorical data and subsequently extend it to data with mixed variables (discrete and continuous).

Like all the methods for finding the association rules of data that include continuous variables, our method cannot find all possible association rules that include continuous variables, as there might be a very large set of possibilities to choose from. One shortcoming of

the method we propose is its inability to perform clustering of all variables (binary, categorical, and continuous) and then determine association rules from these mixed clusters. This shortcoming stems from the fact that we could not identify an appropriate distance function on all types of variables that can be directly connected to the confidence parameter of the association rules (like we did with the Jaccard similarity/distance for the binary variables). We believe that finding such a distance function is a direction worth investigating in future work.

REFERENCES

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami, *Mining association rules between sets of items in large databases*, Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207–216.

[2] Rakesh Agrawal, Ramakrishnan Srikant, et al., *Fast algorithms for mining association rules*, Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.

[3] Yonatan Aumann and Yehuda Lindell, *A statistical theory for quantitative association rules*, Journal of Intelligent Information Systems **20** (2003), no. 3, 255–283.

[4] Sergey Brin, Rajeev Rastogi, and Kyuseok Shim, *Mining optimized gain rules for numeric attributes*, IEEE transactions on knowledge and data engineering **15** (2003), no. 2, 324–338.

[5] Marie Chavent, Vanessa Kuentz, Benoˆıt Liquet, and L Saracco, *Clustofvar: an r package for the clustering of variables*, arXiv preprint arXiv:1112.0295 (2011).

[6] Marie Chavent, Vanessa Kuentz-Simonet, and Jérôme Saracco, *Orthogonal rotation in pcamix*, Advances in Data Analysis and Classification **6** (2012), no. 2, 131–146.

[7] Inderjit S Dhillon, Edward M Marcotte, and Usman Roshan, *Diametrical clustering for identifying anti-correlated gene clusters*, Bioinformatics **19** (2003), no. 13, 1612–1619.

[8] Dheeru Dua and Casey Graff, *UCI machine learning repository*, 2017.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, vol. 1, Springer series in statistics New York, 2001.

[10] G. Gilbert, *Distance between sets*, Nature (1972), no. 239, 174.

[11] Artur Grygorian and Ionut E. Iacob, *A concise proof of the triangle inequality for the jaccard distance*, The College Mathematics Journal **49** (2018), no. 5, 363–365.

[12] Paul Jaccard, *The distribution of the flora in the alpine zone*, New Phytologist **11** (1912), 37–50.

[13] Henk AL Kiers, *Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables*, Psychometrika **56** (1991), no. 2, 197–212.

[14] S. Kosub, *A note on the triangle inequality for the Jaccard distance*, ArXiv e-prints (2016).

[15] M. Levandowsky and D. Winter, *Distance between sets*, Nature (1971), no. 234, 34–35.

[16] Alan H. Lipkus, *A proof of the triangle inequality for the Tanimoto distance*, Journal of Mathematical Chemistry **26** (1999), no. 1, 263–265.

[17] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, Kurt Hornik, et al., *Cluster: cluster analysis basics and extensions*, R package version **1** (2012), no. 2, 56.

[18] Jacinto Mata, José-Luis Alvarez, and José-Cristobal Riquelme, *An evolutionary algorithm to discover numeric association rules*, Proceedings of the 2002 ACM symposium on Applied computing, 2002, pp. 590–594.

[19] Ansaf Salleb-Aouissi, Christel Vrain, and Cyril Nortet, *Quantminer: A genetic algorithm for mining quantitative association rules.*, IJCAI, vol. 7, 2007, pp. 1035–1040.

[20] WS Sarle, *The varclus procedure. sas/stat user's guide*, (1990).

[21] Ramakrishnan Srikant and Rakesh Agrawal, *Mining quantitative association rules in large relational tables*, Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 1996, pp. 1–12.

[22] Kei Takeuchi, Haruo Yanai, and Bishwa Nath Mukherjee, *The foundations of multivariate analysis: a unified approach by means of projection onto linear subspaces*, Wiley New York, 1982.

[23] Evelyne Vigneau and EM Qannari, *Clustering of variables around latent components*, Communications in Statistics-Simulation and Computation **32** (2003), no. 4, 1131–1150.

[24] Ilya E. Vorontsov, Ivan V. Kulakovskiy, and Vsevolod J. Makeev, *Jaccard index based similarity measure to compare transcription factor binding site models*, Algorithms for Molecular Biology **8** (2013), no. 1, 23.

[25] Kunikazu Yoda, Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, and Takeshi Tokuyama, *Computing optimized rectilinear regions for association rules.*, KDD, vol. 97, 1997, pp. 96–103.

APPENDIX A

THE USCENSUS1990 DATASET ATTRIBUTES DESCRIPTION

U.S. DEPARTMENT OF COMMERCE

BUREAU OF CENSUS

*** DATA EXTRACTION SYSTEM ***

DOCUMENTATION OF: FILE CONTENTS

FOR DATA COLLECTION: 'pums901p' - 1990 Decenial Census 1% PUMS - Persons Records

VAR: = Variable Name
TYP: = Variable Type (C = Categorical, N = Numeric Continuous)
DES: = Designation (P = Primary Variable, X = Non-Primary)
LEN: = Length (of the Variable in Characters)
CAT: = Category (of the Variable)

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AAGE	C	X	1		Age Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AANCSTR1	C	X	1		First Ancestry Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AANCSTR2	C	X	1		Second Ancestry Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AAUGMENT	C	X	1		Augmented Pers. See Text Pp. C 5
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ABIRTHPL	C	X	1		Place of Birth
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ACITIZEN	C	X	1		Citizenship Allocation Flag
				0	No

				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ACLASS	C	X	1		Class of Worker Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ADEPART	C	X	1		Time of Departure to Work Allocation Fla
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ADISABL1	C	X	1		Work Limitation Stat. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ADISABL2	C	X	1		Work Prevention Stat. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AENGLISH	C	X	1		Ability to Speak English Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AFERTIL	C	X	1		Chld. Ever Born Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AGE	C	X	2		Age
				00	Less Than 1 Year
				90	90 or More Yrs. Old

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AHISPAN	C	X	1		Detailed Hispanic Origin Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

AHOUR89	C	X	1		Usual Hrs. Worked Per Week in 1989 Alloc
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AHOURS	C	X	1		Hrs. Worked Last Week Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AIMMIGR	C	X	1		Yr. of Entry Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME1	C	X	1		Wages and Salary Inc. Allocation Flag
				0	No
				1	No Derived
				2	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME2	C	X	1		Nonfarm Self Employment Inc. Allocation
				0	No
				1	No Derived
				2	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME3	C	X	1		Farm Self Employment Inc. Allocation Fla
				0	No
				1	No Derived
				2	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME4	C	X	1		Int., Dividend, and Net Rental Inc. Allo
				0	No
				1	No Derived
				2	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME5	C	X	1		Soc. Sec Inc. Allocation Flag
				0	No
				1	No Derived
				2	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

AINCOME6	C	X	1		Pub. Asst. Allocation Flag
				0	No
				1	No Derived
				2	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME7	C	X	1		Ret. Inc. Allocation Flag
				0	No
				1	No Derived
				2	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINCOME8	C	X	1		All Other Inc. Allocation Flag
				0	No
				1	No Derived
				2	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AINDUSTR	C	X	1		Ind. Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ALABOR	C	X	1		Employment Stat. Recode Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ALANG1	C	X	1		Language Other Than English Allocation F
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ALANG2	C	X	1		Language Spoken At Home Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ALSTWRK	C	X	1		Yr. Last Worked Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

AMARITAL	C	X	1		Marital Stat. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AMEANS	C	X	1		Means of Transportation to Work Allocati
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AMIGSTAT	C	X	1		Migration State Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AMOBLLIM	C	X	1		Mobility Limitation Stat. Allocation Fla
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AMOBLTY	C	X	1		Mobility Stat. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ANCSTRY1	C	X	3		Ancestry First Entry See Appendix I Ance
				999	Not Reported

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ANCSTRY2	C	X	3		Ancestry Second Entry See Appendix I Anc
				000	No Secondary Ancestry
				999	Not Reported

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AOCCUP	C	X	1		Occupation Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
APERCARE	C	X	1		Personal Care Limitation Stat. Allocatio
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

APOWST	C	X	1		Place of Work State Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ARACE	C	X	1		Detailed Race Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ARELAT1	C	X	1		Rel. Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ARIDERS	C	X	1		Vehicle Occupancy Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ASCHOOL	C	X	1		School Enrollment Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ASERVPER	C	X	1		Military Per. of Srvc. Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ASEX	C	X	1		Sex Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ATRAVTME	C	X	1		Travel Time to Work Allocation Flag
				0	No
				1	Yes
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AVAIL	C	X	1		Available for Work
				0	N/a Less Than 16 Yrs./at Work/not Lookin
				1	No, Already Has a Job
				2	No, Temply. Ill
				3	No, Other Reasons in School, Etc.

				4	Yes, Could Have Taken a Job

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AVETS1	C	X	1		Military Srvc. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AWKS89	C	X	1		Wks. Worked in 1989 Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AWORK89	C	X	1		Worked Last Yr. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AYEARSCH	C	X	1		Highest Education Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
AYRSSERV	C	X	1		Yrs. of Military Srvc. Allocation Flag
				0	No
				1	Yes

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
CITIZEN	C	X	1		Citizenship
				0	Born in the U.S.
				1	Born in Puerto Rico, Guam, and Outlying
				2	Born Abroad of American Parents
				3	U.S. Citizen by Naturalization
				4	Not a Citizen of the U.s

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
CLASS	C	X	1		Class of Worker
				0	N/a Less Than 16 Yrs. Old/unemp. Who Nev
				1	Emp. of a Private for Profit Company or
				2	Emp. of a Private Not for Profit, Tax Ex
				3	Local Gov. Emp. City, County, Etc.
				4	State Gov. Emp.
				5	Fed. Gov. Emp.
				6	Self Emp. in Own Not Incorp.d Business,
				7	Self Emp. in Own Incorp.d Business, Prof

				8	Working Without Pay in Fam. Bus. or Farm
				9	Unemp., Last Worked in 1984 or Earlier

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
DEPART	C	X	4		Time of Departure for Work Hour and Minu
				0000	N/a Not a Worker or Worker Who Worked At

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
DISABL1	C	X	1		Work Limitation Stat.
				0	N/a Less Than 16 Yrs., and Selected Pers
				1	Yes, Limited in Kind or Amt. of Work
				2	No, Not Limited

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
DISABL2	C	X	1		Work Prevented Stat.
				0	N/a Less Than 16 Yrs., and Selected Pers
				1	Yes, Prevented From Working
				2	No, Not Prevented From Working

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ENGLISH	C	X	1		Ability to Speak English
				0	N/a Less Than 5 Yrs. Old/speaks Only Eng
				1	Very Well
				2	Well
				3	Not Well
				4	Not At All

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
FEB55	C	X	1		Served February 1955 July 1964
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
FERTIL	C	X	2		No. of Chld. Ever Born
				00	N/a Less Than 15 Yrs./male
				01	No Chld.
				02	1 Child
				03	2 Chld.
				04	3 Chld.
				05	4 Chld.
				06	5 Chld.
				07	6 Chld.
				08	7 Chld.
				09	8 Chld.
				10	9 Chld.
				11	10 Chld.

				12	11 Chld.
				13	12 or More Chld.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
HISPANIC	C	X	3		Detailed Hispanic Origin Code See Append
				000	Not Hispanic 006 199
				001	Mexican, Mex Am 210 220
				002	Puerto Rican 261 270
				003	Cuban 271 274
				004	Other Hispanic 200 209, 250 260, 290 401

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
HOUR89	C	X	2		Usual Hrs. Worked Per Week Last Yr. 1989
				00	N/a Less Than 16 Yrs. Old/did Not Work i
				99	99 or More Usual Hrs.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
HOURS	C	X	2		Hrs. Worked Last Week
				00	N/a Less Than 16 Yrs. Old/not At Work/un
				99	99 or More Hrs. Worked Last Week

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
IMMIGR	C	X	2		Yr. of Entry
				00	Born in the U.S.
				01	1987 to 1990
				02	1985 to 1986
				03	1982 to 1984
				04	1980 or 1981
				05	1975 to 1979
				06	1970 to 1974
				07	1965 to 1969
				08	1960 to 1964
				09	1950 to 1959
				10	Before 1950

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME1	C	X	6		Wages or Salary Inc. in 1989
				000000	N/a Less Than 16 Yrs. Old/none
				140000	Topcode
				140001	140001 or More State Median of Topcoded

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME2	C	X	6		Nonfarm Self Employment Inc. in 1989 Sig
				000000	N/a Less Than 16 Yrs./none
				000001	Break Even or $1
				090000	Topcode

				090001	$90, 001 or More State Median of Topcode
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME3	C	X	6		Farm Self Employment Inc. in 1989 Signed
				000000	N/a Less Than 16 Yrs./none
				1	Break Even or $1
				54000	Topcode
				54001	$54001 or More State Median of Topcoded
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME4	C	X	6		Int., Dividends, and Net Rental Inc. in
				000000	N/a Less Than 15 Yrs./none
				1	Break Even or $1
				40000	Topcode
				40001	$40001 or More State Median of Topcoded
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME5	C	X	5		Soc. Sec Inc. in 1989
				00000	N/a Less Than 15 Yrs./none
				17000	Topcode
				17001	17001 or More State Median of Topcoded V
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME6	C	X	5		Pub. Asst. Inc. in 1989
				00000	N/a Less Than 15 Yrs./none
				10000	Topcode
				10001	$10001 or More State Median
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME7	C	X	5		Ret. Inc. in 1989
				00000	N/a Less Than 15 Yrs./none
				30000	Topcode
				30001	$30001 or More State Median of Topcoded
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INCOME8	C	X	5		All Other Inc. in 1989
				00000	N/a Less Than 15 Yrs./none
				20000	Topcode
				20001	$20, 001 or More State Median of Topcode
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
INDUSTRY	C	X	3		Ind. See Appendix I Ind..lst
				000	N/a Less Than 16 Yrs. Old/unemp. Who Nev
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

KOREAN	C	X	1		Served Korean Conflict June 1950 January
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
LANG1	C	X	1		Language Other Than English At Home
				0	N/a Less Than 5 Yrs. Old
				1	Yes, Speaks Another Language
				2	No, Speaks Only English

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
LANG2	C	X	3		Language Spoken At Home See Appendix I L

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
LOOKING	C	X	1		Looking for Work
				0	N/a Less Than 16 Yrs. Old/at Work/did No
				1	Yes
				2	No

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MARITAL	C	X	1		Marital Stat.
				0	Now Married, Except Separated
				1	Widowed
				2	Divorced
				3	Separated
				4	Never Married or Under 15 Yrs. Old

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MAY75880	C	X	1		Served May 1975 to August 1980
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MEANS	C	X	2		Means of Transportation to Work
				00	N/a Not a Worker Not in the Labor Force,
				01	Car, Truck, or Van
				02	Bus or Trolley Bus
				03	Streetcar or Trolley Car
				04	Subway or Elevated
				05	Railroad
				06	Ferryboat
				07	Taxicab
				08	Motorcycle
				09	Bicycle
				10	Walked
				11	Worked At Home

				12	Other Method

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MIGPUMA	C	X	5		Migration Puma State Dependent
				00000	N/a Pers. Less Than 5 Yrs. Old/lived in
				99900	Abroad

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MIGSTATE	C	X	2		Migration State or Foreign Country Code
				00	N/a Pers. Less Than 5 Yrs. Old/lived in
				01	Alabama
				02	Alaska
				04	Arizona
				05	Arkansas
				06	California
				08	Colorado
				09	Connecticut
				10	Delaware
				11	District of Columbia
				12	Florida
				13	Georgia
				15	Hawaii
				16	Idaho
				17	Illinois
				18	Indiana
				19	Iowa
				20	Kansas
				21	Kentucky
				22	Louisiana
				23	Maine
				24	Maryland
				25	Massachusetts
				26	Michigan
				27	Minnesota
				28	Mississippi
				29	Missouri
				30	Montana
				31	Nebraska
				32	Nevada
				33	New Hampshire
				34	New Jersey
				35	New Mexico
				36	New York
				37	North Carolina
				38	North Dakota
				39	Ohio
				40	Oklahoma
				41	Oregon
				42	Pennsylvania
				44	Rhode Island

				45	South Carolina
				46	South Dakota
				47	Tennessee
				48	Texas
				49	Utah
				50	Vermont
				51	Virginia
				53	Washington
				54	West Virginia
				55	Wisconsin
				56	Wyoming
				72	Puerto Rico
				98	Other Abroad in 1985
				99	State Not Identified B Sample

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MILITARY	C	X	1		Military Srvc.
				0	N/a Less Than 16 Yrs. Old
				1	Yes, Now on Active Duty
				2	Yes, on Active Duty in Past, But Not Now
				3	Yes, Srvc. in Reserves or Nat. Guard Onl
				4	No Srvc.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MOBILITY	C	X	1		Mobility Stat. Lived Here on April 1, 19
				0	N/a Less Than 5 Yrs. Old
				1	Yes Same House Nonmovers
				2	No, Different House Movers

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
MOBILLIM	C	X	1		Mobility Limitation
				0	N/a Less Than 15 Yrs./instit. Person, an
				1	Yes, Has a Mobility Limitation
				2	No, Does Not Have a Mobility Limitation

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
OCCUP	C	X	3		Occupation See Appendix I Occup.lst
				000	N/a Less Than 16 Yrs. Old/unemp. Who Nev

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
OTHRSERV	C	X	1		Served Any Other Time
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
PERSCARE	C	X	1		Personal Care Limitation

				0	N/a Less Than 15 Yrs./instit. Person, an
				1	Yes, Has a Personal Care Limitation
				2	No, Does Not Have a Personal Care Limita

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
POB	C	X	3		Place of Birth Appendix I Birth.lst Unit
				001	Alabama
				002	Alaska
				004	Arizona
				005	Arkansas
				006	California
				008	Colorado
				009	Connecticut
				010	Delaware
				011	District of Columbia
				012	Florida
				013	Georgia
				015	Hawaii
				016	Idaho
				017	Illinois
				018	Indiana
				019	Iowa
				020	Kansas
				021	Kentucky
				022	Louisiana
				023	Maine
				024	Maryland
				025	Massachusetts
				026	Michigan
				027	Minnesota
				028	Mississippi
				029	Missouri
				030	Montana
				031	Nebraska
				032	Nevada
				033	New Hampshire
				034	New Jersey
				035	New Mexico
				036	New York
				037	North Carolina
				038	North Dakota
				039	Ohio
				040	Oklahoma
				041	Oregon
				042	Pennsylvania
				044	Rhode Island
				045	South Carolina
				046	South Dakota
				047	Tennessee
				048	Texas
				049	Utah

050	Vermont
051	Virginia
053	Washington
054	West Virginia
055	Wisconsin
056	Wyoming
060	American Samoa
066	Guam
067	Johnston Atoll
069	Northern Mariana Islands
071	Midway Islands
072	Puerto Rico
076	Navassa Island
078	U.S. Virgin Islands
079	Wake Island
081	Baker Island
084	Howland Island
086	Jarvis Island
089	Kingman Reef
095	Palmyra Atoll
096	U.S. Territory, Not Specified
100	Albania
101	Andorra
102	Austria
103	Belgium
104	Bulgaria
105	Czechoslovakia
106	Denmark
107	Faroe Islands
108	Finland
109	France
110	Germany, Not Specified
111	West Germany
112	West Berlin
113	East Berlin
114	East Germany
115	Gibraltar
116	Greece
117	Hungary
118	Iceland
119	Ireland
120	Italy
121	Jan Mayen
122	Liechtenstein
123	Luxembourg
124	Malta
125	Monaco
126	Netherlands
127	Norway
128	Poland
129	Portugal
130	Azores Islands

131	Madeira Islands
132	Romania
133	San Marino
134	Spain
135	Svalbard
136	Sweden
137	Switzerland
138	United Kingdom, Not Specified
139	England
140	Scotland
141	Wales
142	Northern Ireland
143	Guernsey
144	Jersey
145	Isle of Man
146	Vatican City
147	Yugoslavia
148	Europe, Not Specified
149	Central Europe, Not Specified
150	Eastern Europe, Not Specified
151	Lapland, Not Specified
152	Northern Europe, Not Specified
153	Southern Europe, Not Specified
154	Western Europe, Not Specified
180	Union of Soviet Soc.ist Repub.s U.S.
181	Baltic States, Not Specified
182	Estonia
183	Latvia
184	Lithuania
200	Afghanistan
201	Bahrain
202	Bangladesh
203	Bhutan
204	Brunei
205	Burma
206	Cambodia
207	China
208	Cyprus
209	Hong Kong
210	India
211	Indonesia
212	Iran
213	Iraq
214	Israel
215	Japan
216	Jordan
217	Korea, Not Specified
218	South Korea
219	North Korea
220	Kuwait
221	Laos
222	Lebanon

223	Macau
224	Malaysia
225	Maldives
226	Mongolia
227	Nepal
228	Oman
229	Pakistan
230	Paracel Islands
231	Philippines
232	Qatar
233	Saudi Arabia
234	Singapore
235	Spratley Islands
236	Sri Lanka
237	Syria
238	Taiwan
239	Thailand
240	Turkey
241	United Arab Emirates
242	Vietnam
243	Yemen, Peoples Democratic Repub.
244	Yemen Arab Repub.
245	Asia, Not Specified
246	Asia Minor, Not Specified
247	East Asia, Not Specified
248	Gaza Strip
249	Indochina, Not Specified
250	Iraq Saudi Arabia Neutral Zone
251	Mesopotamia, Not Specified
252	Middle East, Not Specified
253	Palestine, Not Specified
254	Persian Gulf States, Not Specified
255	Southeast Asia, Not Specified
256	West Bank
300	Bermuda
301	Canada
302	Greenland
303	St. Pierre and Miquelon
304	North America, Not Specified
310	Belize
311	Costa Rica
312	El Salvador
313	Guatemala
314	Honduras
315	Mexico
316	Nicaragua
317	Panama
318	Central America, Not Specified
330	Anguilla
331	Antigua and Barbuda
332	Aruba
333	Bahamas
334	Barbados
335	British Virgin Islands

336	Cayman Islands
337	Cuba
338	Dominica
339	Dominican Repub.
340	Grenada
341	Guadeloupe
342	Haiti
343	Jamaica
344	Martinique
345	Montserrat
346	Netherlands Antilles
347	St. Barthelemy
348	St. Kitts Nevis
349	St. Lucia
350	St. Vincent and the Grenadines
351	Trinidad and Tobago
352	Turks and Caicos Islands
353	Caribbean, Not Specified
354	Antilles, Not Specified
355	British West Indies, Not Specified
356	Latin America, Not Specified
357	Leeward Islands, Not Specified
358	West Indies, Not Specified
359	Windward Islands, Not Specified
375	Argentina
376	Bolivia
377	Brazil
378	Chile
379	Colombia
380	Ecuador
381	Falkland Islands
382	French Guiana
383	Guyana
384	Paraguay
385	Peru
386	Suriname
387	Uruguay
388	Venezuela
389	South America, Not Specified
400	Algeria
401	Angola
402	Bassas Da India
403	Benin
404	Botswana
405	British Indian Ocean Territory
406	Burkina Faso
407	Burundi
408	Cameroon
409	Cape Verde
410	Central African Repub.
411	Chad
412	Comoros

413 Congo
414 Djibouti
415 Egypt
416 Equatorial Guinea
417 Ethiopia
418 Europa Island
419 Gabon
420 Gambia
421 Ghana
422 Glorioso Islands
423 Guinea
424 Guinea Bissau
425 Ivory Coast
426 Juan De Nova Island
427 Kenya
428 Lesotho
429 Liberia
430 Libya
431 Madagascar

432 Malawi
433 Mali
434 Mauritania
435 Mayotte
436 Morocco
437 Mozambique
438 Namibia
439 Niger
440 Nigeria
441 Reunion
442 Rwanda
443 Sao Tome and Principe
444 Senegal
445 Mauritius
446 Seychelles
447 Sierra Leone
448 Somalia
449 South Africa
450 St. Helena
451 Sudan
452 Swaziland
453 Tanzania
454 Togo
455 Tromelin Island
456 Tunisia
457 Uganda
458 Western Sahara
459 Zaire
460 Zambia
461 Zimbabwe
462 Africa, Not Specified
463 Central Africa, Not Specified
464 Eastern Africa, Not Specified

465	Equatorial Africa, Not Specified
466	French Equatorial Africa, Not Specified
467	French West Africa, Not Specified
468	North Africa, Not Specified
469	Western Africa, Not Specified
470	Southern Africa, Not Specified
500	Ashmore and Cartier Islands
501	Australia
502	Christmas Island, Indian Ocean
503	Clipperton Island
504	Cocos Islands
505	Cook Islands
506	Coral Sea Islands
507	Fiji
508	French Polynesia
509	Kiribati
510	Marshall Islands
511	Micronesia
512	Nauru
513	New Caledonia
514	New Zealand
515	Niue
516	Norfolk Island
517	Palau
518	Papua New Guinea
519	Pitcairn Islands
520	Solomon Islands
521	Tokelau
522	Tonga
523	Tuvalu
524	Vanuatu
525	Wallis and Futuna Islands
526	Western Samoa
527	Oceania, Not Specified
528	Polynesia, Not Specified
529	Melanesia, Not Specified
550	Antarctica
551	Bouvet Island
552	French Southern and Antarctic Lands
553	Heard and Mcdonald Islands
554	At Sea
555	Abroad, Not Specified

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
POVERTY	C	X	3		Pers. Poverty Stat. Recode See Appendix
				000	N/a
				501	501% or More of Poverty Value

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
POWPUMA	C	X	5		Place of Work Puma State Dependent

				00000	N/a Not a Worker Not in the Labor Force,
				99900	Abroad

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
POWSTATE	C	X	2		Place of Work State Appendix I
				00	N/a Not a Worker Not in the Labor Force,
				01	Alabama
				02	Alaska
				04	Arizona
				05	Arkansas
				06	California
				08	Colorado
				09	Connecticut
				10	Delaware
				11	District of Columbia
				12	Florida
				13	Georgia
				15	Hawaii
				16	Idaho
				17	Illinois
				18	Indiana
				19	Iowa
				20	Kansas
				21	Kentucky
				22	Louisiana
				23	Maine
				24	Maryland
				25	Massachusetts
				26	Michigan
				27	Minnesota
				28	Mississippi
				29	Missouri
				30	Montana
				31	Nebraska
				32	Nevada
				33	New Hampshire
				34	New Jersey
				35	New Mexico
				36	New York
				37	North Carolina
				38	North Dakota
				39	Ohio
				40	Oklahoma
				41	Oregon
				42	Pennsylvania
				44	Rhode Island
				45	South Carolina
				46	South Dakota
				47	Tennessee
				48	Texas
				49	Utah

				50	Vermont
				51	Virginia
				53	Washington
				54	West Virginia
				55	Wisconsin
				56	Wyoming
				98	Abroad
				99	State Not Identified

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
PWGT1	C	P	4		Pers. Wgt

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RACE	C	X	3		Recoded Detailed Race Code Appendix C Ra
				001	White 800 869, 971
				002	Black 870 934, 972
				004	Eskimo 935 940, 974
				005	Aleut 941 970, 975
				006	Chinese, Except Taiwanese 605, 976
				007	Taiwanese 606, 607
				008	Filipino 608, 977
				009	Japanese 611, 981
				010	Asian Indian 600, 982
				011	Korean 612, 979
				012	Vietnamese 619, 980
				013	Cambodian 604
				014	Hmong 609
				015	Laotian 613
				016	Thai 618
				017	Bangladeshi 601
				018	Burmese 603
				019	Indonesian 610
				020	Malayan 614
				021	Okinawan 615
				022	Pakistani 616
				023	Sri Lankan 617
				024	All Other Asian 602, 620 652, 985
				025	Hawaiian 653, 654, 978
				026	Samoan 655, 983
				027	Tahitian 656
				028	Tongan 657
				029	Other Polynesian 658, 659
				030	Guamanian 660, 984
				031	Northern Mariana Islander 661, 671, 673
				032	Palauan 663
				033	Other Micronesian 662, 664 670, 672, 674
				034	Fijian 676
				035	Other Melanesian 677 680
				036	Pacific Islander, Not Specified 681 699
				037	Other Race 700 799, 986 999
				301	Alaskan Athabaskan 000, 001, 008, 009, 0

				302	Apache 255 264
				303	Blackfoot 360
				304	Cherokee 416 422, 555 557, 562
				305	Cheyenne 361 363
				306	Chickasaw 436
				307	Chippewa 330 353, 355, 544
				308	Choctaw 226, 228, 404, 434, 520, 559
				309	Comanche 325, 523
				310	Creek 423, 425, 426, 429 432, 449, 540,
				311	Crow 322
				312	Iroquois 405 415
				313	Kiowa 276, 522
				314	Lumbee 464
				315	Navajo 275
				316	Osage 320
				317	Paiute 175 192, 542
				318	Pima 217
				319	Potawatomi 367 374
				320	Pueblo 229 254, 506, 573
				321	Seminole 428, 438 443
				322	Shoshone 195 206, 494, 518
				323	Sioux 282 312, 326, 327
				324	Tlingit 017
				325	Tohono Oodham 218 222
				326	All Other Tribes 002 007, 010 013, 015,
				327	Tribe Not Specified 548, 549, 576 598 Tr

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RAGECHLD	C	X	1		Presence and Age of Own Chld.
				0	N/a Male
				1	With Own Chld. Under 6 Yrs. Only
				2	With Own Chld. 6 to 17 Yrs. Only
				3	With Own Chld. Under 6 Yrs. and 6 to 17
				4	No Own Chld. .incl. Females Under 16 Yrs

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
REARNING	C	X	6		Total Pers. Earnings
				000000	N/a No Earnings
				284000	$284000 Topcode
				284001	State Medians Included

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RECTYPE	C	X	1		Rec. Type
				P	Pers. Record

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RELAT1	C	X	2		Rel. or Not Related or Grp. Qtrs.
				00	Hshldr.

				01	Husband/wife
				02	Son/daughter
				03	Stepson/stepdaughter
				04	Brother/sister
				05	Father/mother
				06	Grandchild
				07	Other Rel.
				08	Roomer/boarder/foster Child
				09	Housemate/roommate
				10	Unmarried Partner
				11	Other Nonrel.
				12	Instit. Person
				13	Other Pers. in Grp. Qtrs.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RELAT2	C	X	1		Detailed Rel. Other Rel.
				0	N/a Gq/not Other Rel.
				1	Son in Law/daughter in Law
				2	Father in Law/mother in Law
				3	Brother in Law/sister in Law
				4	Nephew/niece
				5	Grandparent
				6	Uncle/aunt
				7	Cousin
				8	Other Related by Blood or Marriage
				9	Other Rel.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
REMPLPAR	C	X	3		Employment Stat. of Parents
				000	N/a Not Own Child of Hshldr., and Not Ch
				111	Both Parents At Work 35 or More Hrs.
				112	Father Only At Work 35 or More Hrs.
				113	Mother Only At Work 35 or More Hrs.
				114	Neither Parent At Work 35 or More Hrs.
				121	Father At Work 35 or More Hrs.
				122	Father Not At Work 35 or More Hrs.
				133	Mother At Work 35 or More Hrs.
				134	Mother Not At Work 35 or More Hrs.
				141	Neither Parent in Labor Force
				211	Father At Work 35 or More Hrs.
				212	Father Not At Work 35 or More Hrs.
				213	Father Not in Labor Force
				221	Mother At Work 35 or More Hrs.
				222	Mother Not At Work 35 or More Hrs.
				223	Mother Not in Labor Force

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RIDERS	C	X	1		Vehicle Occupancy
				0	N/a Not a Worker or Worker Whose Means o

				1	Drove Alone
				2	2 People
				3	3 People
				4	4 People
				5	5 People
				6	6 People
				7	7 to 9 People
				8	10 or More People

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RLABOR	C	X	1		Employment Stat. Recode
				0	N/a Less Than 16 Yrs. Old
				1	Civilian Emp., At Work
				2	Civilian Emp., With a Job But Not At Wor
				3	Unemp.
				4	Armed Forces, At Work
				5	Armed Forces, With a Job But Not At Work
				6	Not in Labor Force

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
ROWNCHLD	C	X	1		Own Child See Appendix B, Page 14
				0	Not Own Child
				1	Own Child

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RPINCOME	C	X	6		Total Pers. Inc. Signed
				000000	N/a No Inc.
				401000	Topcode of Total Pers. Income
				401001	State Medians Included

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RPOB	C	X	2		Place of Birth Recode
				10	Born in State of Res.
				21	Northeast
				22	Midwest
				23	South
				24	West
				31	Puerto Rico
				32	American Samoa
				33	Guam
				34	Northern Marianas
				35	Us Virgin Islands
				36	Elsewhere
				40	Born Abroad of American Parents
				51	Naturalized Citizen
				52	Not a Citizen

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

RRELCHLD	C	X	1		Related Child See Appendix B, Page 14
				0	Not Related Child
				1	Related Child
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RSPOUSE	C	X	1		Married, Spouse Present/spouse Absent
				0	N/a Less Than 15 Yrs. Old
				1	Now Married, Spouse Present
				2	Now Married, Spouse Absent
				3	Widowed
				4	Divorced
				5	Separated
				6	Never Married
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
RVETSERV	C	X	2		Veteran Per. of Srvc.
				00	N/a Less Than 16 Yrs. Old, No Active Dut
				01	September 1980 or Later Only
				02	May 1975 to August 1980 Only
				03	May 1975 to August 1980 and September 19
				04	Vietnam Era, No Korean Conflict, No Wwii
				05	Vietnam Era and Korean Conflict, No Wwii
				06	Vietnam Era and Korean Conflict and Wwii
				07	February 1955 to July 1964 Only
				08	Korean Conflict, No Vietnam Era, No Wwii
				09	Korean Conflict and Wwii, No Vietnam Era
				10	Wwii, No Korean Conflict, No Vietnam Era
				11	Other Srvc.
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
SCHOOL	C	X	1		School Enrollment
				0	N/a Less Than 3 Yrs. Old
				1	Not Attending School
				2	Yes, Pub. School, Pub. Coll.
				3	Yes, Private School, Private Coll.
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
SEPT80	C	X	1		Served September 1980 or Later
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
SERIALNO	C	P	7		Hu/gq Pers. Serial No. Unique Within Sta
VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:

SEX	C	X	1		Sex
				0	Male
				1	Female

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
SUBFAM1	C	X	1		Subfam. Rel.
				0	N/a Gq/not in a Subfam.
				1	Husband/wife
				2	Parent in a Parent/child Subfam.
				3	Child in Subfam.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
SUBFAM2	C	X	1		Subfam. Number
				0	N/a Gq/not in a Subfam.
				1	In Subfam. 1
				2	In Subfam. 2
				3	In Subfam. 3

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
TMPABSNT	C	X	1		Temp. Absence From Work
				0	N/a Less Than 16 Yrs. Old/at Work/did No
				1	Yes, on Layoff
				2	Yes, on Vacation, Temp. Illness, Labor D
				3	No

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
TRAVTIME	C	X	2		Travel Time to Work
				00	N/a Not a Worker or Worker Who Worked At
				99	99 Minutes or More to Get to Work

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
VIETNAM	C	X	1		Served Vietnam Era August 1964 April 197
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
WEEK89	C	X	2		Wks. Worked Last Yr. 1989
				00	N/a Less Than 16 Yrs. Old/did Not Work i

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
WORK89	C	X	1		Worked Last Yr. 1989
				0	N/a Less Than 16 Yrs. Old
				1	Worked Last Year
				2	Did Not Work Last Year

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
WORKLWK	C	X	1		Worked Last Week
				0	N/a Less Than 16 Yrs. Old/not At Work/ U
				1	Worked
				2	Did Not Work

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
WWII	C	X	1		Served World War II September 1940 July
				0	Did Not Serve This Per./less Than 16 Yr
				1	Served This Per.

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
YEARSCH	C	X	2		Ed. Attainment
				00	N/a Less Than 3 Yrs. Old
				01	No School Completed
				02	Nursery School
				03	Kindergarten
				04	1st, 2nd, 3rd, or 4th Grade
				05	5th, 6th, 7th, or 8th Grade
				06	9th Grade
				07	10th Grade
				08	11th Grade
				09	12th Grade, No Diploma
				10	High School Graduate, Diploma or Ged
				11	Some Coll., But No Degree
				12	Associate Degree in Coll., Occupational
				13	Associate Degree in Coll., Academic Prog
				14	Bachelors Degree
				15	Masters Degree
				16	Professional Degree
				17	Doctorate Degree

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
YEARWRK	C	X	1		Yr. Last Worked
				0	N/a Less Than 16 Yrs. Old
				1	1990
				2	1989
				3	1988
				4	1985 to 1987
				5	1980 to 1984
				6	1979 or Earlier
				7	Never Worked

VAR:	TYP:	DES:	LEN:	CAT:	VARIABLE/CATEGORY LABEL:
YRSSERV	C	X	2		Yrs. of Active Duty Military Srvc.
				00	N/a Less Than 16 Yrs./no Active Duty Mil
				01	1 Yr. or Less of Srvc.
				50	50 or More Yrs. of Srvc.

APPENDIX B

R CODE

B.1 EXPERIMENT 1

```
#########################################################################
#
# Multiclass Experiment 1: comparison of 1vs1 and DCSVM for various data sets
# using linear SVM.
#
#########################################################################
library(e1071)
library(igraph)

#load functions
source("utils.R")



#generate data: run with one option for a specific data set
#DATA_GEN <- "artificial1"
#DATA_GEN <- "iris"
#DATA_GEN <- "segmentation"
#DATA_GEN <- "letter"
#DATA_GEN <- "heart"
#DATA_GEN <- "wine"
DATA_GEN <- "wine-quality"
#DATA_GEN <- "glass"
#DATA_GEN <- "covertype"
#DATA_GEN <- "svmguide4"
#DATA_GEN <- "vowel"

source("datagen.R")

#create a list of all classes pairs in the training data set
#store in a list

cls.all <- list()
cls.levels <- levels(factor(df.train$class))
cls.n <- length(cls.levels)
idx <- 1
for (i in 1:(cls.n-1)) {
  for (j in (i+1):cls.n) {
    c1 <- cls.levels[i]
    c2 <- cls.levels[j]
    cls.all[[idx]] <- df.train[df.train$class == c1 | df.train$class == c2,]
    cls.all[[idx]]$class <- factor(cls.all[[idx]]$class)
    idx <- idx + 1
  }
}

#create linear svm models for all pairs of classes
svmkernel <- 'linear'
#svmkernel <- 'radial'
```

```
cls.svml <- list()
idx <- 1
for (i in 1:(cls.n-1)) {
  for (j in (i+1):cls.n) {
    c1 <- cls.levels[i]
    c2 <- cls.levels[j]
    #compute weights
    wts <- 100 / table(cls.all[[idx]]$class)
    cls.svml[[idx]] <- svm(class~., data=cls.all[[idx]], kernel=svmkernel, class.weights = wts)
    idx <- idx + 1
  }
}


print('All SVMs created:')
print(Sys.time())

####################################################
#  1vs1 prediction
####################################################
sample <-df.test

p1vs1 <- onevsone(cls.svml, cls.levels, sample)

print(paste('1vs1 prediction results for: "',DATA_GEN,'"',sep = ""))
ans <- table(sample$class == p1vs1)
print(ans)

print(Sys.time())
####################################################
# DCSVM prediction
####################################################

#initialize
info <- initSVMDC(df.train, cls.n, cls.levels, cls.svml, cls.svml)
allpredict <- info$allpredict
svmdc.plan <- createSVMDCplan(cls.n, cls.levels, info$pnodes, astree = T)

print('DCSVM initialization completed:')
print(Sys.time())

#svmdc.plan = make_tree(0,2)
#vertex.id <- 1
#svmdc.plan <- createSVMDCplan(svmdc.plan, NULL,rep(T,cls.n),pnodes,cls.n,cls.levels)

# OPTIONAL: let's print it using a tree-specific layout
# (N.B. you must specify the root node)
co <- layout.reingold.tilford(svmdc.plan, params=list(root=1))
plot.igraph(svmdc.plan, layout=co)

#library(qgraph)
#qgraph(svmdc.plan,edge.labels=T)

#predict and print
```

```
psvmdc <- 0
for (i in 1:nrow(sample)){
  # Pick a single observation for the one-vs-one classifiers to vote on
  candidate = sample[i,]
  vote <- svmdc.predict(svmdc.plan, cls.levels, candidate)

  psvmdc[i] <- vote
}

print(paste('SVMDC prediction results for: "',DATA_GEN,'"',sep = ""))
ans <- table(sample$class == psvmdc)
print(ans)

print(Sys.time())

#compute errors per each class and how classes were mis-classified (the confusion table)
res <- data.frame(orig = sample$class, p1vs1 = p1vs1, psvmdc = psvmdc)
print('Confusion Table 1vs1')
table(res$orig, res$p1vs1, dnn = c("Original", "1vs1"))
print('Confusion Table DCSVM')
table(res$orig, res$psvmdc, dnn = c("Original", "DCSVM"))
```

B.2 THE R CODE FOR COMPUTING CLUSTERING AND ARS

```
##################################################
#
#
##################################################
#set memory limit
memory.limit(6410241024*1024)

#I. READ DATA ###################################################
dsd <- '../data/census/' #data source directory
dsf <- 'USCensus1990raw.data2.csv' #data source file
#dsf <- 'Census_income_clustD59500Q993.csv' #data source file
dsfa <- 'census-varinfo2.csv' #data attributes file

#load data
dfa <- read.csv(paste0(dsd, dsfa), header= T, sep=",")
#df <- read.table(paste0(dsd, dsf), sep = "" , header = F , nrows = 100,
#                 na.strings ="", stringsAsFactors= T)

numvars <- c(13,56,63,64,66:73,91,105)
ccl <- rep('factor', 127)
ccl[numvars] <- 'numeric'
ccl[97] <- 'character'
df <- read.csv(paste0(dsd, dsf), header= F, sep=",", #nrows = 100,
                                na.strings =NA, stringsAsFactors= F,
           colClasses = ccl
                       )
#make REARNING numeric (cannot read numeric from file)
indx <- 97
df[indx] <- lapply(df[indx], function(x) as.numeric(as.character(x)))
```

```
#remove caseid (first column)
toremove <- c(1:12,14:35,38:47,49:53)
df <- df[,-toremove]
header <- c(as.character(dfa[,1]))[-toremove]
colnames(df) <- header


#count missing
sum(is.na(df))

#####################################
#select some categorical vars
cvarsnames <- c(#"Avail",
            #"Citizen",
            #"Class",
            "Disabl1",
            "ENGLISH",
            #"Immigr",
            #"LANG1",
            #"Looking",
            "Marital",
            #"RACE",
            "Sex",
            #"Vietnam",
            #"INDUSTRYCLASS",
            #"WWII",
            "OCCUPCLASS"
            )

set.seed(2020)
dfx <- df[sample(nrow(df), as.integer(nrow(df)*.2)),which(colnames(df) %in% toupper(cvarsnames))]

#convert to binary
#install.packages("dummies")
library(dummies)
dfxb <- dummy.data.frame(dfx)

sum(is.null(dfxb))

#install.packages('proxy')
library(proxy)

#compute all distances between these vectors
d <- dist(t(as.matrix(dfxb)), method = "jaccard")

#hierarchical clustering using Jaccard
groups <- hclust(d,method="complete") #try method = "ward.D", "complete", "single", "average"
#plot dendogram, use hang to ensure that labels fall below tree
plot(groups, hang=-1, main = 'Census clusters', sub = '', xlab = 'Variables')

#####################################
```

```
clusters <- sort(cutree(groups, h = 0.7))
cdf <- data.frame(Variable = names(clusters), Cluster = clusters)
rownames(cdf) <- NULL
library(xtable)
xtable(cdf)

names(clusters[clusters == 1])
names(clusters[clusters == 3])

confidence <- function(df, left, right) {
  rule <- paste0('{', left[1])
  if (length(left) > 1) {
    for (i in 2:length(left)) {
      rule <- paste0(rule, ', ', left[i])
    }
  }
  rule <- paste0(rule, '} => {', right[1])
  if (length(right) > 1) {
    for (i in 2:length(right)) {
      rule <- paste0(rule, ',', right[i])
    }
  }
  rule <- paste0(rule, '}')
  print(rule)

  a = rep(T, nrow(df))
  for (i in 1:length(left)) {
    a = a & df[[left[i]]]
  }
  for (i in 1:length(right)) {
    a = a & df[[right[i]]]
  }
  a = sum(a)
  b = rep(F, nrow(df))
  for (i in 1:length(left)) {
    b = b | df[[left[i]]]
  }
  b = sum(b)

  return (a/b)
}

confidence(dfxb, c("DISABL10","MARITAL4"), c("OCCUPCLASS"))
confidence(dfxb, c("DISABL10"), c("OCCUPCLASS","MARITAL4"))
confidence(dfxb, c("DISABL12"), c("ENGLISH0","SEX1"))
confidence(dfxb, c("ENGLISH0","SEX1"), c("DISABL12"))
confidence(dfxb, c("DISABL10","DISABL12"), c("OCCUPCLASS", "ENGLISH0","SEX1"))
```