CLUSTER ANALYSIS WITH WEIGHTED BINARY VARIABLES

M. K. Kamundi[1], J. M. Kihoro[2], S. M. Mwalili[3] and B. Kiula[4]
[1,2,3]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
[4]Research, Consultancy and Training, Department of ICT Directorate, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
E-mail: kmarykarimi@gmail.com

Abstract
The objective of this study was to discover unique groupings/clusters resulting from performing cluster analysis with weighted binary variables and with binary proximity measures. Cluster analysis techniques were applied to both the simulated binary data and also to the real/survey data that was initially collected to measure the ICT penetration among people in a certain county council in Kenya. For the survey data, only a few indicators (binary variables) were selected for this study. The clustering binary variables used were based on ownership of a Mobile Phone, a Desktop, a Laptop and a Palmtop, for the simulated data; whereas for the survey data they were based on usage of the following: Mobile Data Processing, Mobile Internet, Computer Internet, and Computer Data Processing. For both the simulated and the real/survey data, the names used were fictitious. Ten clusters were identified for the simulated unweighted binary data whereas for the simulated weighted binary data, there were four clusters. Twelve clusters were identified for the real/survey unweighted binary data whereas there were seven clusters for the real weighted binary data. Results of cluster analyses for both the simulated binary data and the real/survey binary data revealed that when the binary variables were weighted very different and unique clusters were formed. Weighting of binary variables was useful in showing that some variables are more important than others and when cluster analysis was performed using the weighted binary variables, unique clusters were formed that portrayed the importance of certain variables.

Key words: Binary variables, binary data, weights, cluster analysis, cluster membership, similarity, distance, dendrogram

1.0 Introduction

Cluster analysis is an exploratory data analysis tool for organizing observed data into meaningful groups or clusters, based on combinations of variables. It is also a tool of discovery revealing associations and structure in data which, though not previously evident, are sensible and useful when discovered.

Cha, *et al.* (2006) proposed weighted binary measurement to improve classification performance based on the comparative study. But in our work we used weighted binary variables but not weighted binary measurements.

Maletta (2007) discussed weighting and its function in statistical analysis of continuous variables and its use in SPSS. Both weighting of cases and weighting of variables are discussed but concentration is only on the weighting of cases. Mentioned also is that in weighting of variables, some of the variables are considered more important than others and hence should be given more weight.

Since Maletta (2007) only considered weighting of cases, we extended his work and explored the effects of using weighted variables and this time weighted binary variables in Cluster Analysis to determine whether membership of clusters will change.

When dealing with binary data there are instance when the totals across binary variables for some binary cases tie. A tie means that the level in the cluster history at which the tie occurred and possibly some of the subsequent levels are not uniquely determined. Cluster Analysis does not seem to have a way to deal with these ties and this is what this paper addresses.

2.0 Methods

2.1 Simulation Model

The simulation model used was the Binomial Distribution, ($X \sim B(n, p)$) (Wiki, 2011) which was used to generate the binary variables with dichotomous outcomes 0 or 1. The number of trials was 50 with varying probabilities for the four binary variables used (mobile phone with probability (0.85), desktop (0.72), laptop (0.46) and palmtop (0.22). For each of the 50 trials (cases) fictitious names of respondents were used. Owning a palmtop was considered most important, followed by laptop, desktop and lastly the mobile phone. The greatest weight was given to palmtop.

A Palmtop or a Personal Data Assistant is a small computer that literally fits in one's palm. It always features a full-fledged operating system. It possesses the capacity of synchronizing with PCs; has expansion slots and communication ports as well. It may have a full physical QWERTY keyboard, and also have the functionality of mobile phones (iPhone, 2010). Since the Palmtop can function as the other 3 items (Mobile phone, Desktop and Laptop) it is the most important among the 4 items.

2.1.1 Real Data

The real data used was survey data obtained from a survey done to measure the ICT penetration in a certain county council in Kenya. Among the indicators used for the survey were:

(i) Use of computer for: calls, e-mail, internet, SMS, word processing and data processing.
(ii) Use of mobile phone for: calls, e-mail, internet, SMS, word processing and data processing.

The data was binary since the questions were asked and answered on a yes/no basis having 1 for 'yes' and 0 for 'no'. Fictitious names of respondents were used for each of the 75 respondents or cases. Indicators or variables of interest were: mobile for data processing and mobile for internet, as well as computer for data processing and computer for internet. The motivation behind the choosing of these four variables is because among them one is more important than the rest, followed by second most important variable and so on.

The capability to use Mobile Technology to carry out Mobile Data Processing was considered most important followed by Mobile Internet, Computer Internet and finally Computer Data Processing. Here the greatest weight was given to Mobile Data Processing.

Basic mobile phones can be used to collect data whereby data is directly fed into a Form already uploaded on to the mobile phone. The mobile phones can be used without wireless network availability because the phone has enough storage capacity to perform basic data analysis. A mobile phone as a data collection tool is an

excellent idea. It is easy to use, reduces time spent – unlike using the paper-based method. Data from the mobile phones is automatically uploaded into the database for analysis (SURE, 2010). For both the simulated and the real data, fictitious names of respondents were used in order to enable easy identification of change of membership of the clusters.

2.3 How the Weights were developed

For the simulated data Principal Components Analysis (PCA) of the simulated data was performed in order to come up with the weights. The weights were selected from the 1st, 2nd, 3rd and 4th principal components.

For the real data sampling on relative frequencies based on the whole population was done. The relative frequencies obtained were then inverted to obtain the weights for the binary variables. The weights were intended to show that some variables are more important than others.

2.4 Cluster Analysis

For the simulated data and the real data, agglomerative hierarchical cluster analysis with Ward's linkage ((Ward, 1963), (Blashfield, 1976), (Kuiper and Fisher, 1975), (Overall et al, 1993) and Squared Euclidean Distance was used to cluster the respondents. Since the study was done on a comparison basis, for consistency the Squared Euclidean Distance was used because of its availability for both the binary data and the interval data.

3.0 Results and Discussion

3.1 Analysis of Simulated Unweighted Data

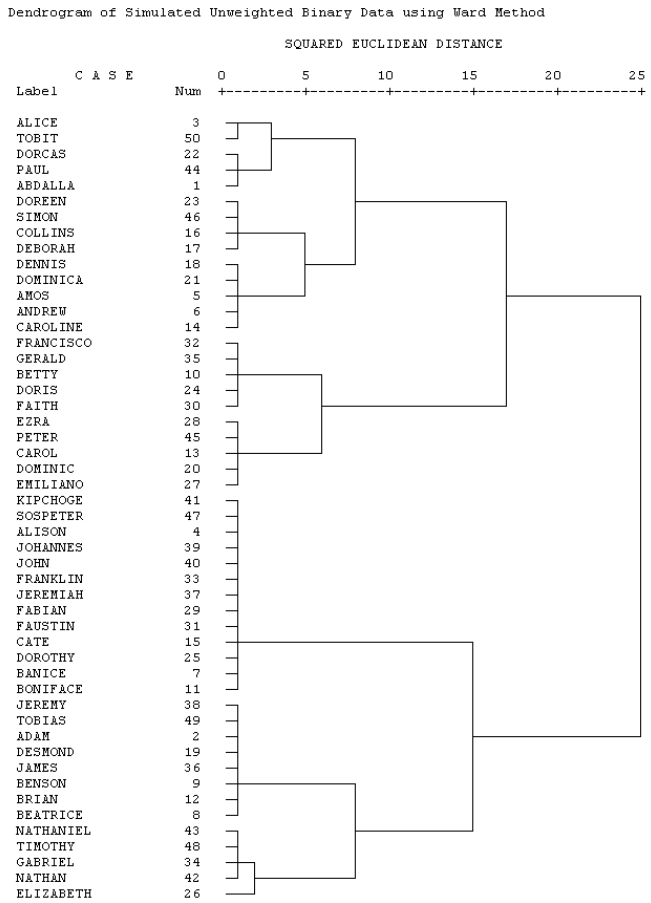

Figure 1: Dendrogram of simulated unweighted data

The dendrogram of simulated unweighted data (Figure 1) contains 10 clusters while those of the simulated weighted data (Figure 2 and Figure 3) contain 4 clusters each.

3.2 Dendrogram of Simulated Unweighted Data

For the dendrogram of simulated unweighted data, the following are the features starting from the bottom to the top of the dendrogram:

- 1st cluster comprises one person only who owns a Desktop.
- 2nd cluster : Both the Left and Right subtrees comprise people who own a Laptop and a Desktop respectively.
- 3rd cluster: Both the Left and Right subtrees comprises people who own a Laptop, a Desktop and a Mobile Phone respectively.
- 4th cluster:Both the Left and Right subtrees comprises people who own a Desktop and a Mobile Phone respectively.
- 5th cluster:Both subtrees comprise people who own all the items – a Palmtop, a Laptop, a Desktop and a Mobile phone respectively.
- 6th cluster: Both subtrees comprise people who own a Palmtop, a Desktop and a Mobile phone respectively.
- 7th cluster: Both subtrees comprise people who own a Laptop and a Mobile Phone respectively.
- 8th cluster:
  Both subtree comprise people who own a Palmtop, a Laptop and a Mobile Phone.
- 9th cluster: Both subtrees comprise people who own a Palmtop and a Mobile Phone respectively.
- 10th cluster: This cluster comprises of people who only own a Mobile Phone.

3.3 Observation from Simulated Unweighted Data Dendrogram

Both the Left and Right subtrees of each cluster contained people with same specific items per cluster. No importance was given to any specific item, hence clustering criteria unknown.

3.4 Analysis of Simulated Weighted Data

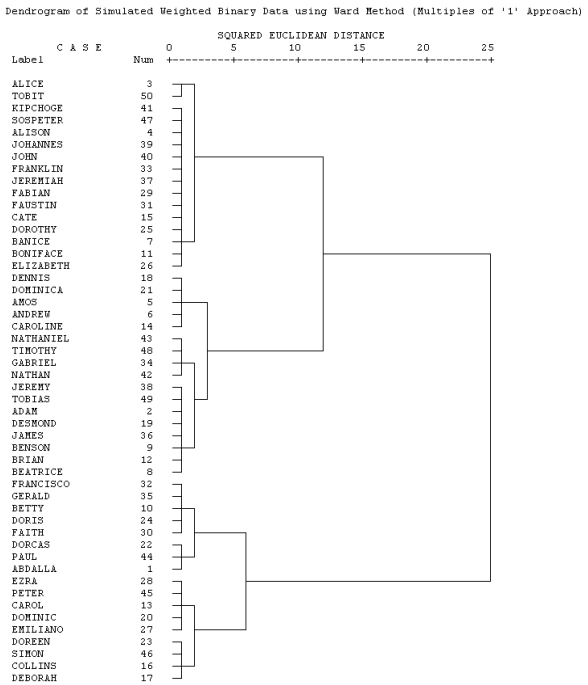

Figure 2: Dendrogram1 of simulated weighted data

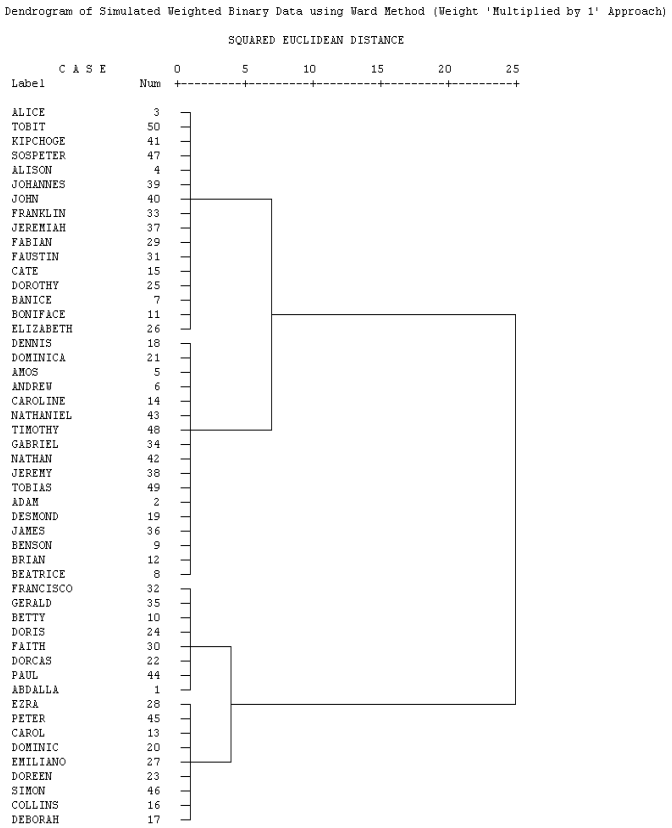

Figure 3: Dendrogram 2 of simulated weighted data

3.5 Dendrograms of Simulated Weighted Data

Figure 2 and Figure 3 are the same. Figure 3 has been cut in such a way that it only reflects the clustering that has been done as a result of **direct** application of weights on the items/variables, whereas Figure 2 reflects clustering of the items/variables using Multiples of '1', where 1 is the default weight of each item/variable. Palmtop was given weight 10; Laptop weight 6; Desktop weight 3 and Mobile Phone weight 2 respectively.

Features of the dendrogram of simulated weighted data (Weight 'Multiplied by 1' Approach) starting from the bottom to the top of the dendrogram:

- 1st cluster: Right subtree comprises of people who own all the 4 items starting with the Palmtop followed by Laptop, Desktop and Mobile phone respectively.
  Left subtree comprises of people who have a Palmtop, a Laptop and a Mobile Phone.
- 2nd cluster: Right subtree comprises of people who own a Palmtop, a Desktop and a Mobile phone respectively.
  Left subtree comprises of people who a Palmtop and a Mobile Phone only.
- 3rd cluster: Right subtree comprises of people who own a Laptop and a Mobile phone only.
  Left subtree comprises of people who own a Laptop and a Desktop only.
- 4th cluster: Right and Left subtrees comprise of people who own a Desktop and a Mobile phone as well as a Mobile Phone only and also a Desktop only.

3.6 Observation from Simulated Weighted Data Dendrogram (Weight 'Multiplied by 1' Approach)

Both the Left and Right subtrees of each cluster contained people with distinct items, with the people clustered according to the weight or importance of the items they own. People who owned Palmtops were given the greatest importance and were placed in the 1st two clusters; followed by people who owned Laptops and they were placed in the 3rd cluster. The 4th cluster contained people who owned Desktop and Mobile Phones which carried lesser and least importance/weight respectively.

By referring to the dendrogram of simulated unweighted data (Figure 1) above, before weighting the data Dennis and Doreen are joined by the same cluster even though they are from consecutive clusters, clusters 7 & 8. Ezra came from cluster 5.

Table 1

DENNIS	1	0	1	0
DOREEN	1	0	1	1
EZRA	1	1	1	1

Now using Table 1 to calculate their Squared Euclidean distances (SED) before weighting:

SED (Dennis,Doreen) = 0 + 0 + 0 + 1 = 1
SED (Dennis,Ezra) = 0 + 1 + 0 + 1 = 2
SED (Doreen,Ezra) = 0 + 1 + 0 + 0 = 1

Based on the above calculation of Squared Euclidean Distances it is evident also from Figure 1 above, that before weighting Dennis and Doreen were placed in the same cluster since they are very similar. But Dennis and Ezra are a bit distant and were placed in different clusters. Since Doreen and Ezra are very similar they should have been placed in the same cluster but were not. This is a case of misclassification which will be addressed when weights are introduced.

Now by referring to the next two dendrograms, dendrogram1 of simulated weighted data (Figure 2) and dendrogram2 of simulated weighted data (Figure 3) where weights have been introduced, Dennis belongs to a different cluster from Doreen and Ezra. Dennis is in cluster 3 whereas Doreen and Ezra are both in cluster 1.

Table 2

DENNIS	2	0	6	0
DOREEN	2	0	6	10
EZRA	2	3	6	10

Using Table 2, their Squared Euclidean Distance after weighting:

SED (Dennis,Doreen) = 100
SED (Dennis,Ezra) = 109
SED (Doreen,Ezra) = 9

Here, weights have been put into consideration portraying the degree of importance of each of the 4 items (Mobile Phone, Desktop, Laptop and Palmtop) with Palmtop carrying the greatest weight. With support from the calculation of Squared Euclidean Distances it is seen from the dendrograms that Dennis was placed in a different cluster from Doreen and Ezra since he was very distant from Doreen and also from Ezra. He was placed in cluster 3. Doreen and Ezra were placed in the same cluster since they were very similar (had the least distance between them). Since Doreen and Ezra both own Palmtops and the Palmtop carries the greatest weight of 10, they were both placed in cluster 1.

In a dendrogram we usually have the left subtree and right subtree. In the dendrogram of weighted binary data the cases/people who own items associated with greatest weight/importance are placed on the right subtree, while those associated with lesser weight/importance are placed on the left subtree.

In the dendrograms of simulated weighted data (Figure 2 and Figure 3), Ezra is placed on the right subtree of cluster 1, while Doreen is placed on the left subtree of the same cluster 1.

3.7 Analysis of Real Unweighted Data

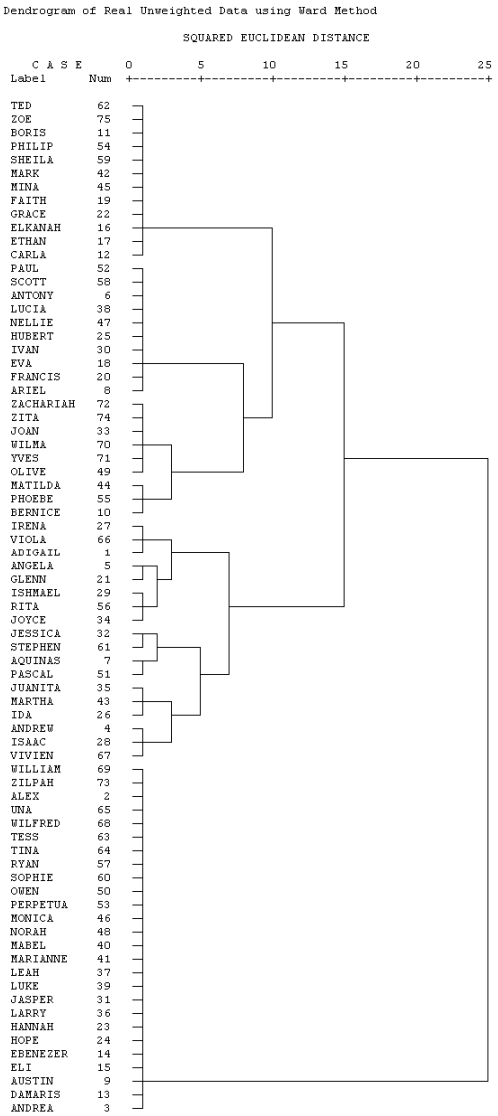

Figure 4: Dendrogram of real unweighted binary data

3.8 Dendrogram of Real Unweighted Data

The dendrogram of real unweighted data (Figure 4) contains 12 clusters while that of the real weighted data (Figure 5) contains 7 clusters. Features of the dendrogram of real unweighted data (Figure 4) starting from the bottom of the dendrogram:

- 1st cluster comprised of people who did not make use of any those services at all.
- 2nd cluster: Right subtree comprises of people who made use of Mobile Data Processing and Computer Internet;
  Left subtree comprises of only one person who made use of Mobile Data Processing, Computer Internet and Computer Data Processing
- 3rd cluster: Both the Right and Left subtrees comprised of people who made use of Mobile Data Processing, Mobile Internet, Computer Internet and Computer Data Processing (all the services).
- 4th cluster: Both the Right and Left subtrees comprised of people who made use of Computer Internet only
- 5th cluster: Both the Right and Left subtrees comprised of people who made use of Mobile Internet and Computer Internet only
- 6th cluster: Right subtree comprised of people who made use of Mobile Data Processing and Mobile Internet.

Left subtree comprised of one person who made use of Mobile Internet only.

- 7th cluster: Both subtrees had people who only made use of Mobile Data Processing only
- 8th cluster: Right subtree comprised of people who made use of Mobile Data Processing and Computer Data Processing only.
  Left subtree had one person who made use of Mobile Data Processing, Mobile Internet and Computer Data Processing.
- 9th cluster: Both subtrees comprised of people who made use of Mobile Internet and Computer Data Processing only
- 10th cluster: Both subtrees comprised of people who made use of Mobile Internet, Computer Internet and Computer Data Processing
- 11th cluster: Both subtrees comprised of people who made use of Computer Internet and Computer Data Processing
- 12th cluster: Both subtrees comprised of people who made use of Computer Data Processing only.

3.9 Observation from dendrogram of Real Unweighted Data

Clustering criteria is unknown.

3.10 Analysis of Real Weighted Data

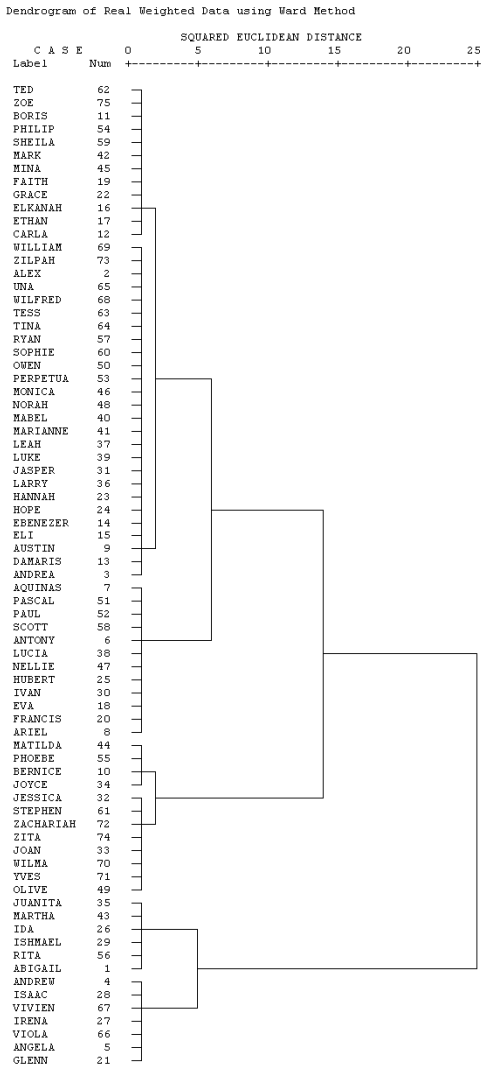

Figure 5: Dendrogram of real weighted binary data

Mobile Data Processing was given weight 4; Mobile Internet weight 3; Computer Internet weight 2 and Computer Data Processing weight 1.

3.11 Dendrogram of Real Weighted Data

Features of the dendrogram of real weighted data (Figure 5) starting from the bottom of the dendrogram:

- 1st cluster: Right subtree comprised of people who made use of Mobile Data Processing and Computer Internet; Mobile Data Processing, Computer Internet and Computer Data Processing.
  Left subtree comprised of people who made use of Mobile Data Processing and Computer Data Processing; and Mobile Data Processing only.
- 2nd cluster: Right subtree comprises of people who made use of all the services (Mobile Data Processing, Mobile Internet,Computer Internet and Computer Data Processing).
  Left subtree comprised of people who made use of Mobile Data Processing and Mobile Internet; Mobile Data Processing, Mobile Internet and Computer Data Processing.
- 3rd cluster: Right subtree comprised of people who made use of Mobile Internet, Computer Internet;
  Left subtree comprised of people who made use of Mobile Internet, Computer Internet and Computer Data Processing.
- 4th cluster: Right subtree comprised of people who made use of Mobile Internet and Computer Data Processing.
  Left subtree comprised of one person who made use of Mobile Internet only
- 5th cluster: Right subtree had people who made use of Computer Internet; Computer Internet and Computer Data Processing;
  Left subtree comprised of people who made use of Computer Internet and Computer Data Processing.
- 6th cluster: Right and Left subtrees comprised of people who made use of none of the services.
- 7th cluster: Both subtrees had people who only made use of Computer Data Processing only.

3.12 Observation from dendrogram of Real Weighted Data

The 1st two clusters give priority to people who make use of Mobile Data Processing since Mobile Data Processing carries the greatest weight/importance. 3rd and 4th clusters had people who made use of Mobile Internet, 5th cluster had people who made use of Computer Internet. 6th cluster had people who made use of none of the services, whereas last cluster, cluster 7 comprised people who made use of Computer Data Processing which had the least weight.

Consider Irena, Viola and Abigail who have been placed together in the 8th cluster in the dendrogram of real unweighted data (Figure 4):

Table 3

IRENA	1	0	0	1
VIOLA	1	0	0	1
ABIGAIL	1	1	0	1

Using Table 3 to calculate the Squared Euclidean Distance (SED) between each of them gives:
SED (Irena,Viola) = 0
SED (Irena,Abigail) = 1
SED (Viola,Abigail) = 1

Irena and Viola are placed in the same cluster since distance between them is zero, hence they are very similar. Abigail was also placed in the same cluster as Irena and Viola since distance from each of them is minimal. They were all (Irena, Viola and Abigail) placed in the 8th cluster in the Real unweighted dendrogram (Figure 4).

But after weights have been introduced in the Real data (refer to Figure 5):

Table 4

IRENA	4	0	0	1
VIOLA	4	0	0	1
ABIGAIL	4	3	0	1

Using Table 4 above, their SED gives:

SED (Irena,Viola) = 0
SED (Irena, Abigail) = $(3)^2$ = 9
SED (Viola,Abigail) = $(3)^2$ = 9

After weighting, Irena and Viola are still very similar since the distance between them is zero hence should be placed in the same cluster. Distance between Irena and Abigail, and also between Viola and Abigail is 9, meaning that Irena is very distant from Abigail and also that Viola is very distant from Abigail. This means that they are not similar and Abigail should not be placed in the same cluster as Irena and Viola. Irena and Viola are hence placed in 1^{st} cluster and Abigail is placed in 2^{nd} cluster.

4.0 Conclusion

Considering both the simulated unweighted data and the real unweighted data, there were circumstances when ties of total across binary variables were evident. When agglomerative hierarchical cluster analysis was performed clusters were formed based on these ties of tallies across variables especially if the tallies were associated with same or matching variables for the cases being clustered. This is usually the case when all variables have equal weight.

But after weighting of the variables of the simulated data and the real data, clusters were formed based on the weights the variables were given. Here, clusters were not formed based on the tallies/totals across binary variables but according to the specific weights assigned to variables. Observations that possess the variables with the greatest weight were clustered on their own; those that possess the variable with the lesser weight only or together with the variables that possess the next lesser weights were grouped in another cluster and so on.

It was evident that when the binary variables were weighted membership of clusters changed and very unique clusters were formed. Membership of clusters changed because different weights were assigned to the binary variables, otherwise if they were all given the same weight, membership could have remained unchanged.

Hence the hypothesis that weighting of binary variables has no effect on membership of clusters is rejected. The conclusion is that this study was successful in proving that when binary variables are weighted the membership of the clusters change resulting in the formation of very unique clusters.
Possible ways of developing the weights could be through using weights or loading generated by Principal Components Analysis. Another way could be to carryout sampling of relative frequencies obtained from the whole population.

Acknowledgements

The Kenya National Bureau of Statistics for providing sponsorship to M. K. Kamundi for the Master of Science in Applied Statistics course.

References

Cha, S. H., Yoon, S. and Tappet, C. C. (2006). "Enhancing Binary Feature Vector Similarities Measures", *Journal of Pattern Recognition Research I*.

Maletta, H. (2007). "Weighting"; Universidad Del Salvador, Buenos Aires, Argentina.

Wiki,(2011). Author: Anonymous; Updated2nd Feb2011 URL: http://wiki.stat.ucla.edu/socr/index.php/Ap_statistics_Curriculum_2007_Distrib_Binomial

iPhone; (2010). iPhone Development; URL: http://palm-freeware.mobi/

SURE (2010). Securing Ugandans' Right to Essential Medicines; Updated: 13 May 2011; URL: http://www.sure.ug/?%26nbsp%3BSuccess_Stories:Mobile_phone_data_collection

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**: pp 236-244.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin* **83**, pp 377-388.

Kuiper, F. K. and Fisher, L. (1975): A Monte Carlo comparison of six clustering procedures. *Biometrics* **31**, pp 777-783.

Overall, J. E., Gibson, J. M. and Novy, D. M. (1993). Population recovery capabilities of 35 cluster analysis methods. *Journal of Clinical Recovery* **49**, pp 459-470.