PROCEEDINGS OF THE INTERNATIONAL STATISTICS CONFERENCE 2011

Contributed Papers

Theme: Statistical Concepts and Methods for the Modern World

28th to 30th December 2011 at Waters Edge, Battaramulla, Sri Lanka

Organised by

Institute of Applied Statistics, Sri Lanka

THE UNIVERSITY OF SYDNEY

School of Mathematics and Statistics, University of Sydney, Australia

Department of Statistics, University of Colombo, Sri Lanka

Editors:

Shelton Peiris (University of Sydney, Australia)
Sarath G. Banneheka (University of Sri Jayewardenepura, Sri Lanka)
Chandima D. Tilakaratne (University of Colombo, Sri Lanka)
Tim B. Swartz (Simon Fraser University, Canada)
S. Ganeshalingam (Massey University, New Zealand)

CLUSTERING DATASETS WITH MIXED TYPES OF ATTRIBUTES: AN APPLICATION TO A NATIONAL YOUTH SURVEY

H. T. Tharanganie

Department of Statistics and Computer Science, University of Sri Jayewardenepura.
{thilakshat@yahoo.com}

ABSTRACT

Clustering is a widely used technique in finding natural groups of data based on some similarity. Most traditional clustering algorithms are limited to handling datasets that contain either continuous or categorical attributes. However, datasets with mixed types of attributes are common in real life applications. Since categorical data has a different structure, the distance functions in the continuous data might not be applicable, and therefore the algorithms developed for continuous data cannot be applied directly to categorical data. Hence, several clustering algorithms have been developed for mixed types of attributes in the literature; however, an approach in SPSS® is accurate and efficient in performance compared to other algorithms. Thus, this study attempts to exemplify the methodologies of SPSS® Two Step Cluster Component which is different from any existing method due to its several desirable features. It extends the model-based distance measure based on the framework of Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) method in order to deal with the datasets with mixed types of attributes. Besides, it handles very large datasets and has the capability to find the optimal number of clusters automatically. The data used is from the Social Policy Analysis and Research Centre (SPARC) of University of Colombo which is the second National Youth Survey conducted with 3000 youths aged 15-29 years in 22 districts of Sri Lanka in 2009. This island wide survey was performed to assess the body of knowledge on attitudes and opinions of young people in Sri Lanka. This dataset consists of 55 mixed types of attributes while two are continuous. Combining both Bayesian Information Criteria (BIC) and distance change, algorithm selected two clusters with around 86% of the cases in the larger cluster. In studying how these two clusters differ, importance of variables in determining the cluster was also examined.

Keywords: Clustering, Mixed Types of Attributes, Categorical Data, Two Step Cluster Analysis, Large Dataset

1. INTRODUCTION

Cluster analysis involves methods that produce classifications from data that are initially unclassified. Clustering methods seek to form 'clusters', 'groups' or 'classes' of individuals such that individuals within a cluster are more 'similar' in some sense than individuals from different clusters (Johnson, 1998). Although many clustering algorithms have been proposed so far, most of those algorithms are designed to find clusters on an assumption that all the attributes are either continuous or categorical. Since datasets with mixed types of attributes

are common in real life applications, the capability to deal with such datasets is undoubtedly important. Although several methods for clustering of both categorical and continuous data have been found in the literature, capabilities of statistical software packages are limited. Thus, understanding the advantages and disadvantages of traditional clustering techniques, SPSS® developed a new method called Two Step (The SPSS Two Step Cluster Component, 2001; Norusis, 2004; Two Step Cluster Analysis, n.d.) to handle mixed types of attributes with a number of enviable characteristics.

This paper is written with the objective of exemplifying the methodologies of SPSS® Two Step Cluster Component using a large dataset with both continuous and categorical variables. This dataset is from the Second National Youth Survey where it is an island wide survey to assess the body of knowledge on attitudes and opinions of young people in Sri Lanka. The Research and Survey Unit of the Social Policy Analysis and Research Centre (SPARC) of the Faculty of Arts, University of Colombo, Sri Lanka in collaboration with International Labour Organization (ILO), United Nations Resident Coordinator's Office (UNRC), Friedrich-Ebert-Stiftung (FES), International Alert (IA), Sri Lanka Youth Parliament, and Sri Lanka Ministry of Youth Affairs conducted this survey with 3000 youths aged 15-29 years in 22 districts of Sri Lanka in 2009. Kilinochchi, Mulativu and Mannar districts in the Northern Province were excluded due to security reasons and in addition, random samples of Vavunia, Batticalo and Trincomalee districts were limited only to accessible areas. Data collection method was the direct face-to-face interaction and a questionnaire was used to collect the data. From a random sample of 3000 respondents, 2997 cases were selected for the analysis due to invalid records. The dataset consists of 55 variables where only 2 variables are continuous. The variables in the dataset represent the following areas: demographic characteristics, youth and culture, language proficiency, attitudes towards social relationships, computer knowledge, politics, ethnic conflicts and violence, educational and job aspirations, and migration.

The overview of the paper is as follows. In Section 2 of this paper, the history of clustering methods and SPSS® Two Step Cluster Component are briefly explained. Section 3 consists of the results of the clustering analyses. All conclusions drawn within the framework of the study are comprehensively discussed in Section 4.

2. MATERIALS AND METHODS

History of Clustering Methods

From the viewpoint of the variables of the target dataset, traditional clustering methods fall into three categories: continuous, categorical and mixed. Most traditional clustering algorithms are limited to handling datasets that contain either continuous or categorical variables. However, some research efforts have been done in mixed types of variables. One algorithm proposed is k-prototype method by Huang (1998). This method is based on the weighted sum of Euclidean distance for continuous variables and a new distance measure on the total mismatches of two data records for categorical variables. However, improper weights may result in biased treatment of different variable types. In Li and Biswas (2002),

the Similarity-Based Agglomerative Clustering (SBAC) algorithm is proposed, which adopts a similarity measure and employs an agglomerative algorithm to construct a dendrogram. However, it is almost impossible to handle very large datasets.

Banfield and Raftery (1993) introduced a model-based distance measure for data with continuous attributes. They derived this measure from a Gaussian mixture model, equivalent to the decrease in log-likelihood resulting from merging two clusters. Based on the framework of Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) method (Zhang *et al.*, 1996), SPSS® Two Step Cluster Component extends this model-based distance measure to situations with both continuous and categorical variables in order to be accurate and efficient in performance compared to other clustering algorithms.

SPSS® Two Step Cluster Component

With over 30 years of experience in statistical software, SPSS® understands the advantages and disadvantages of other statistical methods and applied that knowledge to produce a new method. The SPSS® Two Step Cluster Component handles both continuous and categorical variables by extending the model-based distance measure used by Banfield and Raftery (1993) to situations with mixed type of variables. It utilizes a two-step clustering approach similar to BIRCH (Zhang *et al.*, 1996) which uses the principle of trees and provides the capability to automatically find the optimal number of clusters even in very large datasets. In the first step of the procedure, the records are pre-clustered into many small sub-clusters. Then in the second step, the sub-clusters are clustered from the pre-cluster step into the proper number of clusters automatically. The results gathered from running a simulation are consistently accurate and scalable in performance. The simulation also shows that the automatic procedure of finding the number of clusters works remarkably well and fast.

In SPSS® Two Step Cluster Component, the measure of the relationship between two objects (and also between two clusters) is the log-likelihood distance measure (Two Step Cluster Analysis, n.d.). This probability based distance measure can handle both continuous and categorical variables. In calculating log-likelihood, normal distributions for continuous variables and multinomial distributions for categorical variables are assumed. It is also assumed that the variables are independent of each other, and so are the cases. To find the number of clusters in Two Step cluster procedure, it uses two clustering criterions: Schwarz's Bayesian Information Criterion/BIC (Schwarz, 1978) and Akaike's Information Criterion/AIC (Akaike, 1974). Simulation studies show that combining both clustering criterion values and distance changes works much better than using any one alone. Accordingly, the 'best' cluster solution will have a reasonably large ratio of clustering criterion changes and a large ratio of distance measures.

3. RESULTS AND DISCUSSION

Using the results of the clustering analysis obtained from the SPSS® Two Step Cluster Procedure, the number of clusters, composition of the clusters, and importance of individual variables are examined.

Examining the Number of Clusters

In Two Step Cluster Procedure, SPSS® prints a table of statistics for different numbers of clusters. Table 1 depicts the statistics obtained for this study for different numbers of clusters with Bayesian Information Criterion (BIC) as the clustering criterion. Since the large ratios of BIC changes (1.000) and distance measures (2.547) are occurred for two numbers of clusters, algorithm selected two clusters.

Table 1: Auto-Clustering statistics

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change(a)	Ratio of BIC Changes(b)	Ratio of Distance Measures(c)
1	148537.537			
2	139478.265	-9059.272	1.000	2.547
3	136696.074	-2782.190	.307	1.498
4	135263.418	-1432.656	.158	1.311
5	134473.907	-789.511	.087	1.379
6	134252.172	-221.735	.024	1.119
7	134189.938	-62.234	.007	1.099
8	134248.187	58.249	-.006	1.033
9	134345.109	96.922	-.011	1.216
10	134651.677	306.568	-.034	1.037
11	134992.980	341.303	-.038	1.005
12	135338.841	345.861	-.038	1.161
13	135814.030	475.189	-.052	1.013
14	136299.856	485.826	-.054	1.093
15	136852.879	553.023	-.061	1.011

a The changes are from the previous number of clusters in the table.

b The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Table 2 is identified as the cluster distribution table and it demonstrates the frequency of each cluster. Of the 2997 total cases, 1251 were excluded from the analysis due to missing values on one or more of the variables. Of the 1746 cases assigned to clusters, 1493 were assigned to the first cluster, and 253 to the second. Hence, the larger cluster has around 86% of the clustered cases.

Table 2: Distribution of cases in clusters

		N	% of Combined	% of Total
Cluster	1	1493	85.5%	49.8%
	2	253	14.5%	8.4%
	Combined	1746	100.0%	58.3%
Excluded Cases		1251		41.7%
Total		2997		100.0%

Examining the Composition of Clusters

Once the clusters are formed, it is necessitated to know how they differ. SPSS® offers numerous displays and tables to determine the composition of the clusters. For each continuous variable, a centroids table is obtained. Table 3 displays the mean and standard deviation for the cases in each cluster. Examination of this table provides the evidence that the averages of the two continuous variables in this study are largest for the first cluster. Further examination reveals that the mean 'total monthly family income' for cluster 2 is somewhat lower than the combined mean; however cluster 1 mean is barely higher than the combined mean. Nonetheless, cluster means and standard deviations of the continuous variable 'hours per week spend to watch television' are not largely differing with the combined values.

Table 3: Centroids table for two continuous variables

		Total monthly family income		Hours per week spend to watch television	
		Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	20688.18	13494.863	15.56	9.091
	2	17566.49	10768.198	14.23	9.291
	Combined	20235.84	13178.057	15.37	9.130

For each categorical variable, cross tabulations and bar charts of the distribution of the variable within each cluster are obtained. Table 4 and Figure 2 illustrate these outputs using the particular categorical variable 'Ethnicity'. According to the cluster frequencies of Table 4, cluster 1 is compromised mostly from Sinhalese where majority of cluster 2 contains Tamils

and Moors. Figure 1 shows the percentage of Ethnicity in each of the clusters. It can be seen that Ethnicity distribution in second cluster is largely dissimilar to the overall distribution though the first cluster Ethnicity distribution is fairly similar to the overall distribution. Thus, this figure further clarifies that the respondents' Ethnicity is differ in two clusters.

	Sinhala		Sinhala		Sinhala		Sinhala	
Cluster	Frequency	%	Frequency	%	Frequency	%	Frequency	%
1	1470	99.8	11	6.7	11	10.2	1	50.0
2	3	0.2	152	93.3	97	89.8	1	50.0
Combined	1473	100.0	163	100.0	108	1473	2	100.0

Table 4: Cluster frequency table by Ethnicity

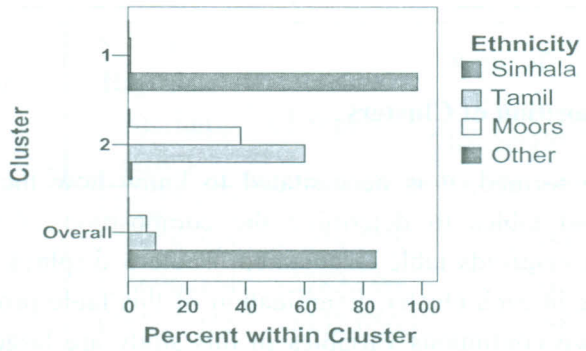

Figure 1: Within-cluster percentage of respondents' Ethnicity

Likewise, for rest of the 52 categorical variables, cross tabulations and bar charts are obtained in order to determine the composition of the clusters. The second cluster with only 14% of the cases has following attributes. Respondents are mostly from Trincomalee, Jaffna, Vavuniya, and Batticalo districts from Northern Province and Eastern Province. Further, most estate sector respondents (95.7%), Tamils (93.3%), Moors (89.8%), Hindus (92.6%), and Muslims (89.0%) are found in the second cluster. Additionally, more than 90% of the respondents in the second cluster are thoroughly capable of speaking, reading and writing Tamil language. Furthermore, around 60% of the second cluster respondents believe that the needs of minority groups in the country have been neglected. Since only these stated attributes have higher percentages for second cluster, they are primarily compromise the composition of that cluster. Conversely, the first cluster has opposite results.

Examining the Importance of Individual Variables

When the cases are clustered, it is vital to discern that how important the different variables are for the formation of the cluster. For continuous variables, SPSS® provides plots of *t* statistics that compare the mean of the variable in the cluster to the overall mean. For a variable to be considered significant, its *t* statistic must exceed the dashed line in either a positive or negative direction. Figure 2 displays such plots for two continuous variables. The

first plot reveals that the average 'total monthly family income' is statistically different for the second cluster, while the second plot demonstrates that the average 'hours per week spend to watch television' is statistically indifferent for two clusters. Thus, it can be concluded that the total monthly family income variable is probably important in distinguishing only the second cluster while the other continuous variable is not important in distinguishing any of the clusters. It is noticeable that these results confirm the trends observed in the Centroids table (Table 3) as only mean 'total monthly family income' for cluster 2 is different with the combined mean.

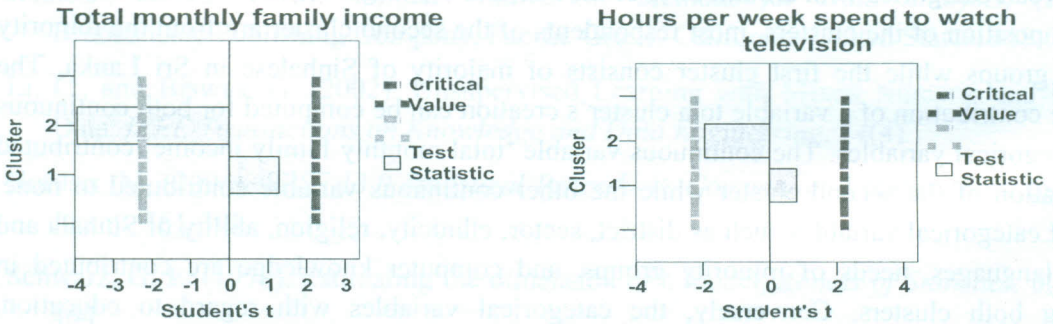

Figure 2: Plots of student's *t* statistics for two continuous variables for each cluster

For categorical variables, SPSS® calculates a Chi-Square value that compares the observed distribution of values of a variable within a cluster to the overall distribution of values. Figure 3 shows the plots of Chi-Square statistics for some categorical variables in this study. The critical value line provides some notion of how dissimilar each cluster is from the average. If the absolute value of the statistic for a cluster is greater than the critical value, the variable is probably important in distinguishing that cluster.

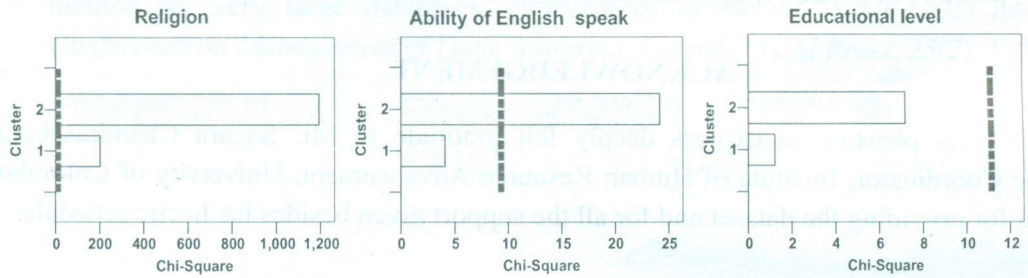

Figure 3: Plots of Chi-Square statistics of some categorical variables

Examinations of these 53 plots of Chi-Square statistics disclose that some Chi-Square statistics are greater than the critical values for both clusters. Hence, these variables are probably important to the formulation for each of the cluster and they are namely district, sector, ethnicity, religion, ability of Sinhala and Tamil languages, needs of minority groups, and computer knowledge of respondents.

On the contrary, average values of the variables are statistically indifferent for both clusters in some plots. It reveals the fact that the variables with regard to education, employment,

social relations, politics, and migration are not important to the formulation of both clusters. However, in few plots, Chi-Square statistics are greater than the critical values for only the second cluster. This exposes that few variables such as ability of English language, marriage, and attitudes towards culture are important to the formulation of only the second cluster.

4. CONCLUSIONS AND RECOMMENDATIONS

Using the SPSS® Two Step Cluster Analysis procedure, the cases are separated into two fairly broad categories with around 86% of the cases of the larger cluster. In the viewpoint of the composition of the clusters, most respondents of the second cluster are from the minority Ethnic groups while the first cluster consists of majority of Sinhalese in Sri Lanka. The relative contribution of a variable to a cluster's creation can be computed for both continuous and categorical variables. The continuous variable 'total monthly family income' contributed the creation of the second cluster while the other continuous variable contributed to none. Several categorical variables such as district, sector, ethnicity, religion, ability of Sinhala and Tamil languages, needs of minority groups, and computer knowledge are contributed in creating both clusters. Conversely, the categorical variables with regard to education, employment, social relations, politics, and migration are not important in any of the clusters. However, few categorical variables such as ability of English language, marriage, and attitudes towards culture are important to the formulation of only the second cluster.

Due to missing values on one or more of the variables, around 42% of the cases were excluded from the analysis in this study. Thus, in order to obtain finer separations within these groups, the complete records with all the schedules filled are highly recommended. In commercial software packages, only the Two Step cluster analysis in SPSS® system makes clustering possible to mixed types of variables; hence researches should focus on this area.

ACKNOWLEDGEMENT

It is with great pleasure to express deeply felt gratitude to Mr. Sagara Chandrasekera, Academic Coordinator, Institute of Human Resource Advancement, University of Colombo, Sri Lanka for providing the dataset and for all the support given besides his hectic schedule.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6): 716-723.

Banfield, J. D., and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49: 803–821.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2: 283-304.

Johnson, D. E. (1998). *Applied Multivariate Methods for Data Analysts*. (2nd ed.) Brooks/Cole Publishing Company, Pacific Grove, California, United States, 319-396.

Li, C., and Biswas, G. (2002). Unsupervised Learning with Mixed Numeric and Nominal Data. *IEEE Transactions on Knowledge and Data Engineering*, *14*(4).

Norusis, M. (2004). *SPSS 13.0 Statistical Procedures Companion*. Prentice Hall, Inc.., New Jersey, 361-391.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2): 461-464.

SPSS Inc., USA. *The SPSS TwoStep Cluster Component*. (2001). Accessed on April 2011 from http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf

TwoStep Cluster Analysis. (n.d.). Accessed on April 2011 from http://www1.uni-hamburg.de/RRZ/Software/SPSS/Algorith.120/twostep_cluster.pdf

Zhang, T., Ramakrishnon, R., and Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal, Canada, ACM Press*, *25*(2): 103-114.