The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 531-538

Revising K-Means Clustering under Semi-Supervision

Myung-Hoe Huh[1], SeongKeun Yi[2], and Yonggoo Lee[3]

Abstract

In k-means clustering, we standardize variables before clustering and iterate two steps: units allocation by Euclidean sense and centroids updating. In applications to DB marketing where clusters are to be used as customer segments with similar consumption behaviors, we frequently acquire additional variables on the customers or the units through marketing campaigns *a posteriori*. Hence we need to modify the clusters originally formed after each campaign. The aim of this study is to propose a revision method of k-means clusters, incorporating added information by weighting clustering variables. We illustrate the proposed method in an empirical case.

Keywords : k-means clustering, customer segmentation, weighting variables, entropy criterion, marketing campaign.

1. Introduction

Suppose that we want to cluster n units into k groups using p attributes $X_1, ..., X_p$. In DB marketing applications, statisticians are asked to segment large number of customers into groups of similar consumption behaviors, which are captured into clustering variables. Due to its efficiency in handling large data sets, k-means clustering is regarded as the standard method in data mining (Giudici, 2003).

In k-means clustering, clustering variables $X_1, ..., X_p$ need to be standardized into $Z_1, ..., Z_p$ which have mean 0 and standard deviation 1. As result, the (squared) distance between units i and i' is defined as

$$d_E^2(i,i') = \sum_{j=1}^{p} (z_{ij} - z_{i'j})^2 , \quad i, i' = 1, \cdots, n$$

in Euclidean sense. This setting implies *a priori* equality in importance of all clustering variables. Although such implicit assumption is inevitable in the absence of any other

[1] Professor, Dept. of Statistics, Korea University. Anam-Dong 5, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

[2] Assistant Professor, Dept. of Business Administration, Sungshin Women's University, Dongseon-Dong 3, Sungbuk-Gu, Seoul 136-742, Korea.

[3] Professor, Dept. of Statistics, Chung-Ang University. Huksuk-Dong 221, Seoul 156-756, Korea.

information on individual unit's response propensity, we need to allow unequal importance of clustering variables upon the availability of unit response variable Y (or $y_1, \cdots, y_n$ for n units). In DB marketing which is the contextual background of this study, customer's reaction for various marketing campaigns may serve as the response variable Y.

Instantly, one may consider supervised learning on Y by $X_1, ..., X_p$. But it is not quite useful in DB marketing when a series of campaigns are launched to the panel of customers in a relatively short period, since separate models for each Y requires time-consuming pilot studies. Certainly, multipurpose grouping of units or unsupervised learning meets marketer's minimum needs even though it does not accomodate campaign results.

The aim of this study is to develop a revising method of k-means clustering reflecting recent campaign results for the next campaign. Thus the proposed method can be called by semi-supervised k-means clustering, since it updates existing k-means clustering using interim unit responses.

Optimal weighting scheme of variables for cluster analysis first appeared in DeSarbo, Carrol, Clark and Green (1984) but never a top performer. For weighting of variables in k-means clustering as unsupervised learning, Makarenkov and Legendre (2001) adopted a derivative based method, which differs from our method in the aim.

2. Algorithm for Weighting Variables

Suppose that we assign weights $w_1, \cdots, w_p$ to $Z_1, ..., Z_p$, standardized versions of $X_1, ..., X_p$, where

$$w_1, \cdots, w_p > 0 \quad \text{and} \quad \sum_{j=1}^{p} w_j = p.$$

Then, the (squared) distance between units i and i' is modified to

$$d^2_W(i, i') = \sum_{j=1}^{p} w_j (z_{ij} - z_{i'j})^2$$

which may be called weighted Euclidean distance. We propose to derive weights for variables that best accomodate available unit responses represented by Y. To simplify prescriptions, we assume that Y is categorical with nominal codes from 1 to q. Our method can be written as the following algorithm:

Step 1: K-means clustering is executed on n units of p measurements to obtain k groups of self-similar units.

Step 2: At the time Y is acquired, 1) randomly generate $w_1, \cdots, w_p$, by placing $p-1$ random points on the interval of length p, 2) weight variables by these numbers. In other words, multiply $Z_1, ..., Z_p$ by $w_1^{0.5}, \cdots, w_p^{0.5}$, and 3) execute another k-means clustering using Step 1's clusters as initial grouping.

Step 3: Evaluate the collective entropy Ent with respect to Y existing in clusters:

$$Ent = \sum_{g=1}^{k} (n_g/n) Ent_g \ ,$$

where n_g is the size of cluster $g\,(=1, \dots, k)$, n is the whole sample size, and Ent_g is the entropy measured at the cluster g. Specifically,

$$Ent_g = -\sum_{l=1}^{q} r_{gl} \log_2 r_{gl}$$

for $(r_{g1}, \cdots, r_{gq})$denotes the composition of cluster $g = 1, \cdots, k$. Ent_g and Ent varies with $w_1, \cdots, w_p$. If Ent is smaller than any former ones, then retain current weights and clusters. Otherwise, discard current weights and clusters.

Step 4: Repeat Step 2 and Step 3 for a sufficiently large number of times.

Briefly we will demonstrate our method with well-known Fisher's iris data. Merits of the proposed method will be illustrated with a more realistic case in the next section.

In Fisher's iris data, there are four measurements $X_1, \dots, X_4$ (sepal length, sepal width, petal length, petal width) for each of 150 iris flowers. We will apply our method with k = 3 with the known species information Y (=1:setosa, 2:versicolor, 3:virginica).

1) K-means clustering yields the cluster centroids on standardized scale:

	Z1	Z2	Z3	Z4	Size
Cluster 1	1.03	0.01	0.94	0.97	55
Cluster 2	-0.17	-0.97	0.26	0.17	46
Cluster 3	-1.00	0.90	-1.30	-1.25	49

2) Among 400 random trials of various weights $w_1, \cdots, w_p$, we selected the best k-means clusters inherited from ordinary k-means clustering that has the smallest entropy with respect to the species (=1,2,3). The best weights for $X_1, \dots, X_4$ are

$$w_1 = 0.1754,\ w_2 = 0.2624,\ w_3 = 2.3063,\ w_4 = 1.2559.$$

Note that X_3 and X_4 are weighted more compared to two other variables. On weighted scale, cluster centroids are

	Z1(W)	Z2(W)	Z3(W)	Z4(W)	Size
Cluster 1	0.40	-0.06	1.56	1.24	49
Cluster 2	0.03	-0.37	0.44	0.18	51
Cluster 3	-0.42	0.44	-1.98	-1.40	50

3) Now compare two cross-classified tables between clusters and the species:

Ordinary K-means Clustering → Weighted K-Means Clustering

	Seto	Versi	Virgi	Seto	Versi	Virgi
Cluster 1	0	13	42	0	2	47
Cluster 2	1	37	8	0	48	3
Cluster 3	49	0	0	50	0	0

Observe that the weighted clustering has only five miss-classifications, reduced from 22 of the ordinary k-means clustering. This makes sense since the weighted clustering is obtained using additional information, i.e., species tag in this case. From ordinary clustering to weighted clustering, cluster memberships of 19 units have changed. The following matrix shows the switching pattern.

	Weighted Clustering		
Ordinary	C1	C2	C3
Cluster 1	43	12	0
Cluster 2	6	39	1
Cluster 3	0	0	49

To see the stability of our Monte Carlo algorithm, we repeated the same process as above, and obtained the following results.

- Best weights for $X_1, \ldots, X_4$: w_1 = 0.0199, w_2 = 0.2316, w_3 = 2.7409, w_4 = 1.0076.

- Cross-classified tables between clusters and the species:

	Ordinary K-means Clustering			→	Weighted K-Means Clustering		
	Seto	Versi	Virgi		Seto	Versi	Virgi
Cluster 1	0	13	42		0	3	47
Cluster 2	1	37	8		0	47	2
Cluster 3	49	0	0		50	0	0

- Switching pattern from the ordinary K-means clustering to the weighted clustering:

	Weighted Clustering		
Ordinary	C1	C2	C3
Cluster 1	44	11	0
Cluster 2	7	38	1
Cluster 3	0	0	49

Thus, we see that the cross-classified table and switching pattern matrix are very similar to the former one, even though derived variable weights appear somewhat different.

In the next section, we apply our method to customer segmentation problem in which campaign results are acquired in sequence.

3. Telecommunication Case

The case is for telecommunication company, Telco. Telco collects and manages its customer's socio-demographic profiles, calling behaviors, billing records and campaign results. Here we analyze Telco data given in SPSS Version 12.0.

In market segmentation research using clustering method, the choice of variables, or segmentation bases should be by theory or by experience on the domain of application. The selection of bases is based on the researcher's judgement of their relevance for the type segments being sought and substantive segmentation at hand (Milligan, 1995; Wedel and Kamakura, 2000).

SPSS Telco data consists of 1,000 sample records, each of which contains 42 variables. Among various variables, we selected five variables on calling behaviors for clustering purpose. Clustering variables are X1: long distance last month, X2: toll free last month, X3: equipment rental last month, X4: calling card last month, X5: wireless last month. The data also contains a series of campaign results such as the "Caller ID" and the "Call Waiting".

We executed k-means clustering with k = 5 and obtained the following as five cluster centroids (on standardized scale):

	Z1	Z2	Z3	Z4	Z5	Size
Cluster 1	2.53	0.12	-0.32	1.56	-0.28	69
Cluster 2	-0.83	9.45	-0.75	-0.30	-0.59	1
Cluster 3	-0.33	-0.64	-0.21	-0.55	-0.54	503
Cluster 4	-0.11	0.45	1.55	0.28	1.52	190
Cluster 5	0.05	0.92	-0.71	0.49	0.01	237

We observe that Cluster 2 consists of only one customer with very large toll-free calls (Z2). Thus Cluster 2 customer is apparently an outlier. Thus, hereafter, we exclude that unit and rerun the analysis with the number of clusters k = 4. Firstly, we obtained the following as four cluster centroids (on newly standardized scale):

	Z1	Z2	Z3	Z4	Z5	Size
Cluster 1	2.28	0.02	-0.33	1.53	-0.33	80
Cluster 2	-0.10	0.48	1.58	0.28	1.51	187
Cluster 3	0.02	1.05	-0.70	0.42	0.04	233
Cluster 4	-0.34	-0.67	-0.21	-0.55	-0.53	499

Secondly, we modified k-means clustering just after "Caller ID" campaign, which resulted in 518 failures(F) and 481 successes(S). Cluster centroids produced by 400 Monte Carlo trials are (on weighted scale):

	Z1(W)	Z2(W)	Z3(W)	Z4(W)	Z5(W)	Size	F	S
Cluster 1	0.18	-1.43	0.05	0.66	-0.02	145	97	48
Cluster 2	0.18	3.27	0.20	0.64	1.65	115	4	111
Cluster 3	-0.02	1.12	0.00	0.09	0.09	343	79	264
Cluster 4	-0.10	-1.39	-0.08	-0.50	-0.55	396	338	58

At this time, it turns out that Cluster 2 has the highest customer value, followed by Cluster 3, Cluster 1, Cluster 4 in descending order of values. The weights assigned to clustering variables are

$$w_1 = 0.0734,\ w_2 = 3.1355,\ w_3 = 0.1706,\ w_4 = 0.5159,\ w_5 = 1.0986.$$

We note that X2 (=toll free) is heavily weighted while X1 (=long distance) and X3 (=equipment) are almost negligible. As result, internal cluster characteristics changed quite a bit, also did cluster memberships of customers. Following matrix shows the pattern of membership switching.

		New Cluster			
		1	2	3	4
Old Cluster	1	49	11	19	1
	2	33	54	92	8
	3	4	50	179	0
	4	59	0	53	387

We see that cluster memberships of 330 units on off-diagonals have changed, while 669 units on diagonals remained within the same cluster.

Thirdly, we re-modified k-means clustering again after "Call Waiting" campaign, which resulted in 514 failures(F) and 485 successes(S). Cluster centroids produced by another 400 Monte Carlo trials are (on weighted scale):

	Z1(W)	Z2(W)	Z3(W)	Z4(W)	Z5(W)	Size	F1F2	F1S2	S1F2	S1S2
Cluster 1	0.88	-1.28	0.02	0.70	-0.07	110	68	8	11	23
Cluster 2	0.34	2.36	0.16	0.46	2.06	147	4	2	5	136
Cluster 3	-0.06	1.16	-0.04	0.09	-0.19	307	30	44	29	204
Cluster 4	-0.29	-1.29	-0.03	-0.39	-0.55	435	323	39	44	29

Here, F1F2 denotes two consecutive failures, F1S2 denotes F(=fail) followed by S(=success), and so on. The weights assigned to clustering variables are

$$w_1 = 0.5479,\ w_2 = 2.68731,\ w_3 = 0.0431,\ w_4 = 0.4443,\ w_5 = 1.2774.$$

Note that X2(=toll free) lost part of its weight but still is most important and that X1(=long distance) gained the weight notably. As result, cluster characteristics changed a little again. However, the value order of clusters does not change: Cluster 2 should be most highly valued, followed by Cluster 3, Cluster 1, Cluster 4. Following matrix shows the pattern of membership switching.

		New Cluster			
		1	2	3	4
Old Cluster	1	97	1	0	47
	2	0	101	14	0
	3	3	45	293	2
	4	10	0	0	386

We observe that cluster memberships of 122 units on off-diagonals have changed, while 877 units on diagonals remained within the same cluster. In this way, sequential revision of the k-means clustering can be done.

Now, let us evaluate the merits due to cluster updating. If market researcher uses the original k-means clusters for the second campaign "Call Waiting" as well as for the first campaign "Caller ID", then it turns out that the response rates are 0.87 for Cluster 3, 0.78 for Cluster 2, 0.40 for Cluster 1, 0.21 for Cluster 4. See Table 1. On the other hand, if he/she updates the original k-means clusters after the first campaign "Caller ID" for the next campaign, then it turns out that the response rates are 0.95 for Cluster 2, 0.81 for Cluster 3, 0.28 for Cluster 1, 0.14 for Cluster 4. Thus the ranges of response rates are 0.66 (=0.87-0.21) for the first case and 0.81 (=0.95-0.14) for the second case. The larger range of the two supports the cluster revision plan.

Table 1. Comparison of original k-means clustering vs. revised k-means clustering in turnouts of "Call Waiting" campaign

Original K-Means Clustering					Revised K-Means Clustering				
	F	S	Total	Response %		F	S	Total	Response %
Cluster 1	48	32	80	0.40	Cluster 1	104	41	145	0.28
Cluster 2	41	146	187	0.78	Cluster 2	6	109	115	0.95
Cluster 3	31	202	233	0.87	Cluster 3	64	279	343	0.81
Cluster 4	394	105	499	0.21	Cluster 4	340	56	396	0.14
Total	514	485	999	0.49	Total	514	485	999	0.49

4. Concluding Remarks

For unsupervised k-means clustering, Makarenkov and Legendre (2001) adopted a derivative-based method for determining optimal weights for clustering variables. In contrast, our algorithm of Section 2 determines optimal weights for clustering variables simply by Monte Carlo method. Perhaps, one may develop more efficient algorithm that meets the purpose of this study. Still, with modern environment of personal computers, such brute force computing is affordable. For instance, it took about fifteen minutes of CPU time for processing a case analysis of Section 3 with our notebook computer (Intel Pentium 4 Mobile CPU 1.4 GHz; 512 Mega Bytes RAM).

Setting the number of clusters k could be an issue here. In Section 3, we set k to five for SPSS Telco data, by the reproducibility evaluation method of Huh and Lee (2004), which is based on Hubert-Arabie's corrected Rand index for clustering rules from partitioned data sets.

The proposed method for revising k-means clustering by weighting variables is somewhere between unsupervised and supervised learning. Toward more supervised direction, competing alternative is radial basis functional network or RBFN (Ripley, 1996). There are however, two clear differences: 1) RBFN normally assumes that all explanatory variables (corresponding to clustering variables of this study) are scaled equally, and 2) RBFN is directly aimed at individual prediction rather than clustering of individual units.

Someone may be concerned about the magnitude of fluctuations or over-fitting problem in variable weights during a series of cluster updating processes. The problem can be fixed, we guess, by introducing the concept of learning rate that controls the degree of the change in variable weights. It needs further study.

References

[1] DeSarbo, W.S., Carrol, J.D., and Clark, L.A., and Green, P.E. (1984). "Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables," *Psychometrika*, 49. 57-78.

[2] Giudici, P. (2003). *Applied Data Mining.* Wiley. (Section 4.2)

[3] Huh, M.H, and Lee, Y. (2004). "Reproducibility assessment of k-means clustering and applications," *Korean Journal of Applied Statistics*, 17. 135-144. (Written in Korean)

[4] Makarenkov, V. and Legendre, P. (2001). "Optimal variable weighting for ultrametric and additive trees and k-means partitioning: methods and software," *Journal of Classification*, 18. 245-271.

[5] Milligan, G. W. (1995). "Issues in applied classification," *CSNA Newsletter*, 36-38.

[6] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press. (Section 4.2)

[7] Wedel, M. and Kamakura, W. (2000). *Market Segmentation: Concept and Methodological Foundations,* 2nd Ed. Kluwer Academic Publishers. (Chapter 5)

[Received January 2005, Accepted July 2005]