Latent class models for clustering: A comparison with K-means

Jay Magidson
Statistical Innovations Inc.
Jeroen K. Vermunt
Tilburg University

Recent developments in latent class (LC) analysis and associated software to include continuous variables offer a model-based alternative to more traditional clustering approaches such as K-means. In this paper, the authors compare these two approaches using data simulated from a setting where true group membership is known. The authors choose a setting favourable to K-means by simulating data according to the assumptions made in both discriminant analysis (DISC) and K-means clustering. Since the information on true group membership is used in DISC but not in clustering approaches in general, the authors use the results obtained from DISC as a gold standard in determining an upper bound on the best possible outcome that might be expected from a clustering technique. The results indicate that LC substantially outperforms the K-means technique. A truly surprising result is that the LC performance is so good that it is virtually indistinguishable from the performance of DISC.

Introduction

In the last decade, there has been a renewed interest in latent class (LC) analysis to perform cluster analysis. Such a use of LC analysis has been referred to as the mixture likelihood approach to clustering (McLachlan & Basford 1988; Everitt 1993), model-based clustering (Banfield & Raftery 1993; Bensmail, *et al.* 1997; Fraley & Raftery 1998a, 1998b), mixture-model clustering (Jorgensen & Hunt 1996; McLachlan, *et al.* 1999), Bayesian classification (Cheeseman & Stutz 1995), unsupervised learning (McLachlan & Peel 1996), and latent class cluster analysis (Vermunt & Magidson 2000, 2002).

Probably the most important facilitating reason for the increased popularity of LC analysis as a statistical tool for cluster analysis is that high-speed computers now make these computationally intensive methods practically applicable. Several software packages are available for estimating LC cluster models.

An important difference between standard cluster analysis techniques and LC clustering is that the latter is a model-based approach. This means that a statistical model is postulated for the population from which the data sample is obtained. More precisely, it is assumed that a mixture of underlying probability distributions generates the data.

When using the maximum likelihood method for parameter estimation, the clustering problem involves maximizing a log-likelihood function. This is similar to standard non-hierarchical cluster techniques such as K-means clustering, in which the allocation of objects to clusters should be optimal according to some criteria. These criteria typically involve minimizing the within-cluster variation or, equivalently, maximizing the between-cluster variation. An advantage of using a statistical model is that the choice of the cluster criterion is less arbitrary and the approach includes rigorous statistical tests.

LC clustering is very flexible as both simple and complicated distributional forms can be used for the observed variables within clusters. As in any statistical model, restrictions can be imposed on the parameters to obtain more parsimony, and formal tests can be used to check their validity. Another advantage of the model-based clustering approach is that no decisions have to be made about the scaling of the observed variables. For instance, when working with normal distributions with unknown variances, the results will be the same irrespective of whether the variables are normalized. This is very differ-

ent from standard non-hierarchical cluster methods like K-means, where scaling is always an issue. Other advantages are that it is relatively easy to deal with variables of mixed measurement levels (different scale types) and that there are more formal criteria to make decisions about the number of clusters and other model features.

In the marketing research field, LC clustering is sometimes referred to as latent discriminant analysis (Dillon & Mulani 1989) because of the similarity to the statistical methods used in discriminant analysis (DISC) as well as logistic regression (LR). However, an important difference is that in discriminant and logistic regression modelling, group (cluster) membership is assumed to be known and observed in the data while in LC clustering it is unknown (latent) and, therefore, unobservable.

In this paper, we use a simple simulated data set to compare the traditional K-means clustering algorithm as implemented in the SPSS (KMEANS) and SAS (FASTCLUS) procedures with the latent class mixture modeling approach as implemented in the Latent GOLD (Vermunt & Magidson 2000) package. We use discriminant analysis as the gold standard in evaluating the performance of both approaches.

Research design

For simplicity, we consider the case of two continuous variables that are normally distributed within each of two clusters (i.e., two populations, two classes) with variances the same within each class. These assumptions are made in DISC. We also assume that within each class, the variables are independent of each other (local independence), a prerequisite of the K-means approach. Formally, we generate two samples according to the following specifications:

Within the kth population, $y = (y_1, y_2)$ is characterized by the bivariate normal density $f_k(y \mid \mu_k, \sigma_k, \rho_k)$. We set $\mu_1 = (3,4)$, $\mu_2 = (7,1)$, $\sigma_1 = \sigma_2 = (2,1)$, and $\rho_1 = \rho_2 = 0$. Samples of size N_1=200 and N_2=100 were drawn at random from these populations with results given in Table 1.

Within each population, DISC assumes that y follows a bivariate normal distribution with common variances and covariances; assumptions met by our generated data. Under these assumptions, it also follows that the probability of group membership satisfies the LR model (see appendix). Hence, use of both DISC and LR are justified here and will be used as standards by which to evaluate the results of K-means and LC clustering.

Table 1
Design parameters and sample statistics for generated data

	Class 1		Class 2	
Parameter	Population Value	Sample Estimate (N=200)	Population Value	Sample Estimate (N=100)
Mean (Y_1)	3	3.09	7	6.89
Mean (Y_2)	4	3.95	1	1.28
Std dev. (Y_1)	1	1.06	1	0.99
Std dev. (Y_2)	2	2.19	2	1.75
Correlation (Y_1,Y_2)	0	0.05	0	–0.08

In real-world clustering applications, supervised learning techniques such as DISC and LR cannot be used since information on group membership would not be available. Unsupervised learning techniques such as LC cluster and K-means need to be used when group membership is unobservable. For classifying cases in this application, the information on true group membership will be ignored when using the unsupervised techniques. Hence, the unsupervised techniques can be expected to perform somewhat worse than the supervised techniques. We will judge the performance of the unsupervised methods by observing how good the results are in comparison to the supervised techniques.

Results obtained from the supervised learning techniques

We show in the appendix how the classification performance for DISC, LR, and LC cluster analysis can be evaluated by estimating and comparing the associated equiprobability (EP) lines $y_2 = \alpha' + \beta' y_1$ for each technique. Cases for which (y_1, y_2) satisfies this equation are equally likely to belong to population 1 or 2 (i.e., the posteri-

or probability of belonging to population 1 and 2 are both 0.5). Cases falling to the left of the line are predicted to be in population 1, those to the right are predicted to be in population 2. Figures 1a and 1b show the EP-lines estimated from the DISC and LR analyses, respectively. (See the appendix for the derivation of these lines.)

Figure 1a shows that only one case from population 1 falls to the right of the EP line obtained from DISC and so is incorrectly predicted to belong to population 2. Similarly, three cases from population 2 fall to the left of the line and are incorrectly predicted to belong to population 1. Comparison of Figures 1a and 1b shows that the line estimated from LR provides slightly worse discrimination of the two populations and misclassifies one additional case from the population 1.

Figure 1a
Equi-probability line estimated using discriminant analysis

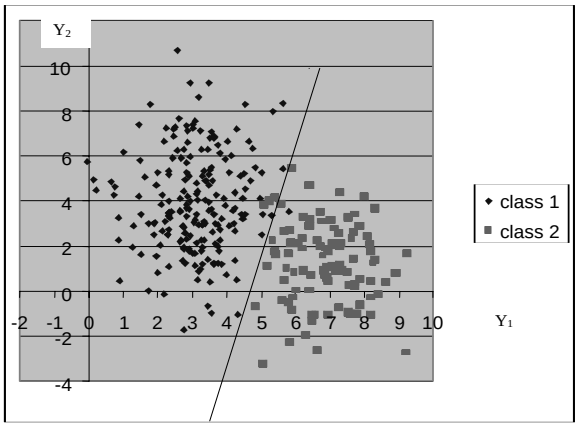
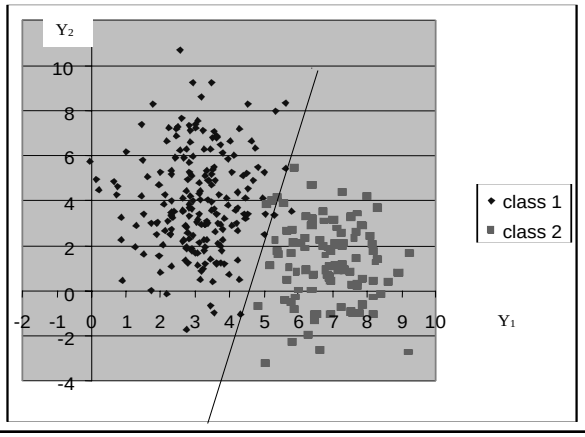


Figure 1b
Equi-probability line estimated using logistic regression



DISC correctly classifies 199 of the 200 population 1 cases and 97 of the 100 cases from population 2, resulting in an overall misclassification rate of 1.3%. This compares to an overall rate of 1.7% obtained using LR. The results of these supervised techniques are summarized in Table 2.

Table 2
Results from the supervised learning techniques

		Misclassified	
		Discriminant	Logistic
Population	Total	Analysis	Regression
Class 1	200	1	2
Class 2	100	3	3
Total	300	4 (1.3%)	5 (1.7%)

Results obtained from the unsupervised learning techniques

For the unsupervised techniques, true class membership is assumed to be unknown and must be predicted. Because LC clustering is model-based, basic model parameters are estimated for class size, as well as means and variances of the variables within each class. These estimates were found to be very close to the corresponding sample values which were used for the DISC analysis (Table 3). As a result of the closeness of these estimates, the resulting EP-line obtained from LC clustering turns out to be virtually indistinguishable from the corresponding EP-line obtained from DISC (Figure 1a). A comparison of the slope and intercept of the EP-lines obtained from DISC, LR, and LC clustering is given in Table 4. (The formula showing how these EP-line parameters are obtained as a function of the basic parameter estimates is given in the appendix.)

Classification based on the modal posterior membership probability estimated in the LC clustering procedure resulted in only three of the cluster 1 cases and one of the cluster 2 cases being misclassified — an overall misclassification rate of 1.3%, a performance equal to the gold standard obtained by DISC.

Table 3
Comparison of LC Cluster estimates from the corresponding sample values used by DISC

	Class 1		Class 2	
Parameter	DISC	LC Cluster	DISC	LC Cluster
Class size	200	200.8	100	99.2
Mean (Y_1)	3.09	3.11	6.89	6.88
Mean (Y_2)	3.95	3.94	1.28	1.28
Var (Y_1)	1.09	1.14	1.09	1.14
Var (Y_2)	4.23	4.23	4.23	4.23
Correlation (Y_1,Y_2)	0.05	0*	–0.08	0*

* Restricted to zero according to the local independence assumption made by K-means clustering

Table 4
Comparison of the parameter estimates of the equi-probability line $Y_2 = \alpha' + \beta' Y_1$ with the population values

	α'	β'
Population	–25.09	5.33
DISC	–25.06	5.33
LR	–25.98	5.54
LC Cluster	–24.90	5.28

Table 5 shows that the overall misclassification rate for K-means is 8%, substantially worse than LC clustering. However, the results of K-means, unlike those of LC clustering and DISC are highly dependent upon the scaling of the variables (results from DISC and LC clustering are invariant of linear transformations of the variables).

Table 5
Results from unsupervised learning techniques

Population	Total	LC Cluster	K-means	K-means with standardized variables
Class 1	200	1	18	10
Class 2	100	3	6	5
Total	300	4 (1.3%)	24 (8.0%)	15 (5.0%)

Therefore, we performed a second K-means analysis after standardizing the variables Y_1 and Y_2 to Z-scores. Table 5 shows that this second analysis yields an improved misclassification rate of 5%, but one that remains significantly worse than that of LC clustering.

If the variables had been standardized in a way that the within cluster variance of Y_1 and Y_2 were equated, the K-means technique would have performed on par with LC cluster and DISC. This type of standardization, however, is not possible when cluster membership is unknown. The overall Z-score standardization served to bring the within class variances of Y_1 and Y_2 closer together, but the within class variance of Y_2 remained significantly larger than that of Y_1.

Another factor to consider when comparing LC clustering with K-means is that for unsupervised techniques, the number of clusters is also unknown. In the case of K-means, the researcher must determine the number of classes without relying on formal diagnostic statistics since none are available. In LC modelling, various statistics are available that can assist in choosing one model over another.

Table 6 shows the results of estimating six different LC cluster models to these data. The first four models vary the number of classes between 1 and 4. The BIC statistic correctly selects the standard two-class model as best. The remaining LC models are variations of the two-class

Table 6
Results from estimation of several LC cluster models

Model	Log Likelihood	BIC	Number of Model Parameters
1-Cluster equal	–1333	2689	4
2-Cluster equal	–1256	2552*	7
3-Cluster equal	–1251	2558	10
4-Cluster equal	–1250	2574	13
2-Cluster unequal	–1252	2555	9
2-Cluster equal + corr	–1256	2557	8

* This model is preferred according to the BIC criterion (lowest BIC value)

model that relax certain assumptions of the model. Specifically, the fifth model includes two additional variance parameters (one for each variable) to relax the DISC assumption of equal variances within each class. The final model adds a correlation parameter, relaxing the K-means assumption of local independence. The BIC statistic correctly selects the standard two-class model as best among all of these LC cluster models.

Summary and conclusion

Our results suggest that LC performs as well as discriminant analysis and substantially better than K-means for this type of clustering application. More generally, for traditional clustering applications where the true classifications are unknown, the LC approach offers several advantages over K-means. These include:

1. Probability-based classification. While K-means uses an *ad hoc* approach for classification, the LC approach allows cases to be classified into clusters using model-based posterior membership probabilities estimated by maximum likelihood (ML) methods. This approach also yields ML estimates for misclassification rates. (The expected misclassification rate estimated by the standard two-class LC cluster model in our example was 0.9%, comparable to the actual misclassification rate of 1.3% that was achieved by LC cluster and DISC. These misclassification rates were substantially better than that obtained using K-means.) Another advantage of assigning a probability to cluster membership is that it prevents biasing the estimated cluster-specific means; that is, in LC analysis an individual contributes to the means of cluster k with a weight equal to the posterior membership probability for cluster k. In K-means, this weight is either 0 or 1, which is incorrect in the case of misclassification. Such misclassification biases the cluster means, which in turn may cause additional misclassifications.

2. Determination of number of clusters. K-means provides no assistance in determining the number of clusters. In contrast, LC clustering provides various diagnostics such as the BIC statistic, which can be useful in determining the number of clusters.

3. No need to standardize variables. Before performing K-means clustering, analysts must standardize variables to have equal variance to avoid obtaining clusters that are dominated by variables having the most variation. Such standardization does not completely solve the problems associated with scale differences since the clusters are unknown and so it is not possible to perform a within-cluster standardization. In contrast, the LC clustering solution is invariant of linear transformations on the variables, so standardization of variables is not necessary.

Additional advantages relate to use of various extensions of the standard LC cluster models that are possible, such as:

1. More general structures can be used for the cluster-specific multivariate normal distributions. More precisely, the (unrealistic) assumption of equal variances and the assumption of zero correlations can be relaxed.

2. LC models can be estimated where the number of latent variables is increased instead of the number of clusters. These LC factor models have been found to outperform the traditional LC cluster models in several applications (Magidson & Vermunt 2001).

3. Inclusion of variables of mixed scale types. K-Means clustering is limited to interval scale quantitative variables. In contrast, extended LC models can be estimated in situations where the variables are of different scale types. Variables may be continuous, categorical (nominal or ordinal), or counts or any combination of these (Vermunt & Magidson 2000, 2002). If all variables are categorical, one obtains a traditional LC model (Goodman 1974).

4. Inclusion of demographics and other exogenous variables. A common practice following a K-means clustering is to use discriminant analysis to describe differences between the clusters on one or more exogenous variables. In contrast, the LC cluster model can be easily extended to include exogenous variables (covariates). This allows both classification and cluster description to be performed simultaneously using a single uniform ML estimation algorithm.

Limitations of this study

The fact that DISC outperformed LR in this study does not mean that DISC should be preferred to LR. DISC obtains maximum likelihood (ML) estimates under

bivariate normality, an assumption that holds true for the data analyzed in this study. LR, on the other hand, obtains conditional ML estimates without reliance on any specific distributional structure on y. Hence, in other simulated settings where the bivariate normal assumption does not hold, LR might outperform DISC since the DISC assumption would be violated.

Another limitation of the study was that we simulated data from a single model representing the most favourable case for K-means clustering. We showed that even in such a situation, LC clustering outperforms K-means. When data are simulated from less favourable situations for K-means, such as unequal within-cluster variances and local dependencies, the differences between K-means and LC clustering are much larger (Magidson & Vermunt 2002).

References

Anderson, T.W. (1958) *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Banfield, J.D. & A.E. Raftery (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49: 803–21.

Bensmail, H., G. Celeux, A.E. Raftery, & C.P. Robert (1997) Inference in model based clustering. *Statistics and Computing*, 7: 1–10.

Cheeseman, P. & J. Stutz (1995) Bayesian classification (Autoclass): Theory and results. U.M.Fayyad,

Piatetsky-Shapiro, G., P. Smyth, & R.Uthurusamy (eds.) *Advances in knowledge discovery and data mining*. Menlo Park: The AAAI Press.

Dillon, W.R. & N. Mulani (1989) LADI: A latent discriminant model for analyzing marketing research data. *Journal of Marketing Research*, 26: 15–29.

Everitt, B.S. (1993) *Cluster analysis*. London: Edward Arnold.

Fraley, C. & A.E. Raftery (1998a) MCLUST: Software for model-based cluster and discriminant analysis. Department of Statistics, University of Washington: Technical Report No. 342.

——— (1998b) How many clusters? Which clustering method? Answers via model-based cluster analysis. Department of Statistics, University of Washington: Technical Report no. 329.

Goodman, L.A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61: 215–31.

Jorgensen, M. & L. Hunt (1996) Mixture model clustering of data sets with categorical and continuous variables. Proceedings of the Conference ISIS '96, Australia: 375–84.

Magidson J. & J.K. Vermunt (2001) Latent class factor and cluster models, bi-plots and related graphical displays. In *Sociological Methodology 2001*. Cambridge: Blackwell: 223–64.

——— (2002) Latent class modeling as a probabilistic extension of K-means clustering. *Quirk's Marketing Research Review*, March.

McLachlan, G.J. & K.E. Basford (1988) *Mixture models: Inference and application to clustering*. New York: Marcel Dekker.

McLachlan, G.J. & D. Peel (1996) An algorithm for unsupervised learning via normal mixture models. In D.L. Dowe, K.B.Korb, & J.J.Oliver (eds.), *Information, Statistics and Induction in Science*. Singapore: World Scientific Publishing: 354–63.

McLachlan, G.J., D. Peel, K.E. Basford, & P. Adams (1999) The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, 4(2).

Vermunt, J.K. & J. Magidson (2000) *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.

——— (2002) Latent class cluster analysis. In J.A. Hagenaars & A.L. McCutcheon (eds.), *Advances in Latent Class Analysis*. Cambridge University Press.

About the authors

Jay Magidson is founder and president of Statistical Innovations Inc., a Boston-based consulting, training, and software development firm. He designed the SPSS CHAID and GOLDMineR® programs and is co-developer (with Jeroen Vermunt) of the Latent GOLD® program.

Jeroen K. Vermunt is associate professor in the Methodology and Statistics Department of the Faculty of Social and Behavioral Sciences at Tilburg University in the Netherlands. His research interests include latent class modelling, event history analysis, and categorical data analysis.

Appendix

Case 1: Cluster proportions known. The sample sizes ($n_1 = 200$, $n_2 = 100$) were drawn proportional to the known population sizes. Hence, the overall probability of belonging to population 1 is twice that of belonging to population 2 : $\pi_1 = 2/3$, $\pi_2 = 1/3$, $\pi_1/\pi_2 = 2$. In the absence of information of the value of y for any given observation, the probability of belonging to population 1 is given by the a priori probability $\pi_1 = 2/3$.

For a given observation $y = (y_1, y_2)$, following Anderson (1958), the posterior probabilities of belonging to populations 1 and 2 can be defined using Bayes theorem as:

$$\pi_{1|y} = \frac{\pi_1 f_1(y)}{\pi_1 f_1(y) + \pi_2 f_2(y)} \quad \text{and} \quad \pi_{2|y} = 1 - \pi_{1|y} = \frac{\pi_2 f_2(y)}{\pi_1 f_1(y) + \pi_2 f_2(y)}$$

Thus, the posterior odds of belonging to population 1 is:

$$\frac{\pi_{1|y}}{\pi_{2|y}} = \frac{\pi_1}{\pi_2} \frac{f_1(y)}{f_2(y)} \qquad (1)$$

Eq. (1) states that for any given y, the posterior odds can be computed by multiplying the *a priori* odds by the ratio of the densities evaluated at y. Hence, the ratio of the densities serves as a Bayes factor. Further, when the densities are BVN with equal variances and equal covariance, the posterior odds follows a linear logistic regression:

$$\ln\left(\frac{\pi_{1|y}}{\pi_{2|y}}\right) = \alpha + \beta_1 y_1 + \beta_2 y_2$$

Under the current assumptions, since $\rho = 0$, we have:

$$\alpha = \ln(\pi_1 / \pi_2) - 1/2\left[\frac{(\mu_{11}^2 - \mu_{21}^2)}{\sigma_1^2} + \frac{(\mu_{12}^2 - \mu_{22}^2)}{\sigma_2^2}\right], \qquad (2.1)$$

$$\beta_1 = (\mu_{11} - \mu_{21}) / \sigma_1^2 \quad \text{and} \quad \beta_2 = (\mu_{12} - \mu_{22}) / \sigma_2^2 \qquad (2.2)$$

In this case, an observation will be assigned to population 1 when $\pi_1 > .5$, which occurs when

$\ln\left(\frac{\pi_{1|y}}{\pi_{2|y}}\right) = \alpha + \beta_1 y_1 + \beta_2 y_2 > 0$, or when $y_2 > \alpha' + \beta' y_1$.

where

$$\alpha' = -(\alpha / \beta_2) \text{ and } \beta' = -(\beta_1 / \beta_2) \qquad (3)$$

In LC cluster analysis, ML parameter estimates are obtained for the quantities on the right hand side of eqs. (2.1) and (2.2). Substitution of these estimates in eqs. (2.1) and (2.2) yield ML estimates for α, β_1 and β_2 which can be used in eq. (3) to obtain the parameters of the EP-line α', and β'.

Appendix

Case 2: Unknown population size

More generally, when the population proportions are not known and the sample is not drawn proportional to the population, the uniform prior can be used for the *a priori* probabilities in which case the *a priori* odds π_1/π_2 equals 1, and α reduces to:

$$\alpha = -1/2[\frac{(\mu_{11}{}^2 - \mu_{21}{}^2)}{\sigma_1^2} + \frac{(\mu_{12}{}^2 - \mu_{22}{}^2)}{\sigma_2^2}] \quad (4)$$

In the case that $\sigma_1^2 = \sigma_2^2$, eq (1) reduces further to:

$$\frac{\pi_{1|y}}{\pi_{2|y}} = \frac{f_1(y)}{f_2(y)} = \exp(\alpha + \beta_1 y_1 + \beta_2 y_2) \quad (5)$$

where $\alpha = 1/2(\mu_{21}{}^2 - \mu_{11}{}^2 - \mu_{12}{}^2 + \mu_{22}{}^2)$

$$\beta_1 = \mu_{11} - \mu_{21}$$

and

$$\beta_2 = \mu_{12} - \mu_{22}$$