

# Simultaneous supervised and unsupervised classification modeling for assessing cluster analysis and improving results interpretability

Mario Fordellone and Maurizio Vichi

**Abstract** In the unsupervised classification field, the unknown number of clusters and the lack of assessment and interpretability of the final partition by means of inferential tools, denote important limitations that could negatively influence the reliability of the final results. In this work, we propose to combine unsupervised classification with supervised methods in order to enhance the assessment and interpretation of the obtained partition. In particular, the approach consists in combining of the clustering method  $k$ -means (KM) with logistic regression (LR) modeling to have an algorithm that allows an evaluation of the partition identified through KM, to assess the correct number of clusters, and to verify the selection of the most important variables. An application on real data is presented to better clarify the utility of the proposed approach.

**Key words:** Supervised Classification, Unsupervised Classification, Assessing Clustering

## 1 Introduction

In unsupervised classification techniques, clusters of homogeneous objects are detected by means of a set of features measured (observed) on a set of objects without knowing the membership of objects to clusters. In these applications, the aim is to discover the heterogeneous structure of the data. In unsupervised classification models, the principal approaches of cluster analysis [6] are: connectivity-based clustering better known as hierarchical clustering, centroid-based clustering,

---

Mario Fordellone  
Sapienza, University of Rome e-mail: mario.fordellone@uniroma1.it

Maurizio Vichi  
Sapienza, University of Rome e-mail: maurizio.vichi@uniroma1.it

distribution-based clustering, density-based clustering, and many other parametric and non-parametric techniques [7].

Conversely, supervised classification is based on the idea of forecasting the membership of new objects (output) based on a set of features (inputs) measured on a training set of objects for which the membership to clusters is known. Therefore, in these applications, the aim is to generalize a function or mapping from inputs to outputs which can then be used speculatively to generate an output for previously unseen inputs [4] [8]. Usually, a subsample (training), which is representative of specific groups, is selected and then this model is used as reference for the classification of new (unobserved) other objects. Training sets are selected based on the knowledge of the user. In supervised classification models we have artificial neural networks, naive Bayes classifiers, nearest neighbor algorithm naive, decision trees, logistic regression, generalized linear models, and many other parametric and non-parametric techniques.

In unsupervised classification, we have important issues that could drastically influence results: (i) an unknown number of clusters, (ii) an absence of variable selection that most contribute to clustering, and (iii) a final assessment of clusters [3]. In other words, all the decisions taken to address the study can lead to different results and each single decision becomes crucial for the aim of our study and needs to be tested.

In this work, we propose an algorithm based on the use of supervised classification modeling. In particular, our approach consists in the simultaneous application of  $k$ -means (KM) [9] and logistic regression (LR) [1] modeling. We will prove that, by using LR, we have effective inferential tools for choosing the number of clusters, selecting the most important variables for the clustering, and assessing the quality of clusters.

The paper is structured as follows: in section 2 we present our proposal for the simultaneous application of unsupervised and supervised classification modeling, in section 3 we show an application on real data and finally, in section 4 we try to give some suggestions and concluding remarks on the work.

## 2 Proposal

In unsupervised classification modeling, we are not interested in prediction because we do not have an associated response variable  $y$  like in supervised classification modeling. Therefore, this paper proposes to simultaneously apply unsupervised (i.e., KM) and supervised classification (i.e., LR) approaches, where the latter aims to evaluate and to improve the former with the additional data structure information. We will call this approach  $k$ -means-logistic regression (KM-LR). In particular, KM-LR is composed of the following principal steps:

Given the  $n \times J$  data matrix  $\mathbf{X}$ , for  $K = 2, \dots, Kmax$ , where  $Kmax$  is the maximum number of clusters the researcher thinks the data might have, the algorithm works as follows:

**Algorithm 1** KM-LR algorithm

---

```

1: for  $k = 2$  to  $Kmax$  do
2:   K-means step
3:   Randomly initialize the membership matrix  $\mathbf{U}$ ;
4:   Compute the centroids matrix by  $\tilde{\mathbf{X}} = (\mathbf{U}^T \mathbf{U})^{-1} \mathbf{U}^T \mathbf{X}$ ;
5:   Minimize the objective function  $\|\mathbf{X} - \mathbf{U}\tilde{\mathbf{X}}\|^2$  with respect to the membership matrix  $\mathbf{U}$ ;
6:   Update the centroids matrix  $\tilde{\mathbf{X}}_n = (\mathbf{U}_n^T \mathbf{U}_n)^{-1} \mathbf{U}_n^T \mathbf{X}$  given the new assignment matrix  $\mathbf{U}_n$ ;
7:   if  $\|\mathbf{X} - \mathbf{U}_n \tilde{\mathbf{X}}_n\|^2 > \omega$ ;
      $\tilde{\mathbf{X}} = \tilde{\mathbf{X}}_n$ ,  $\mathbf{U} = \mathbf{U}_n$ , repeat steps 5-6;
     else
       exit loop; obtain the  $g_k$  categorical cluster vector;
     end if
8:   end if
9:   Multinomial logistic regression step
10:  LR is estimated on  $g_k$ , with explanatory variables  $\mathbf{X}$ , for estimating the probabilities for its  $k - 1$  response
11:  categories  $\pi_k(\mathbf{x})$ , and to estimate the probabilities for its baseline category  $\pi_0(\mathbf{x})$ ;
12:  if some LR estimated coefficient is not 5% statistically significant;
     remove the corresponding variables from the matrix  $\mathbf{X}$ ;
     repeat steps 2-10;
     end if
13: end for

```

---

At the end, we obtain  $Kmax - 1$  identified partitions (with a different number of clusters  $k$ ), together with a reduced set of statistically significant variables and a set of inferential tools to assess the quality of the partition. The best partition (with the optimal number of clusters  $k$ ) is identified in correspondence of the largest increase of a  $\chi^2$ -test computed on the partitions obtained by KM and LR. In this way, through the analysis of the LR results (e.g., explained variance, parameters significance, residual variance), we have an evaluation of the partition obtained by KM. In fact, a good performance of the LR model on the response variable derived by the KM outcome means that the variables included in the model provide a good explanation for the group structure in the data. Moreover, through the LR coefficients analysis, we can see which variables contribute the most to identifying the group structure and to what extent they do so (by analyzing statistical significance, value estimates, and signs of coefficients).

Note that the algorithm monotonically decreases the loss function, or at least does not increase it. However, it does not guarantee to stop at the global minimum of the loss function. For this reason, it is recommended to use of a large number of randomly started runs to find the best solution. The predictive accuracy of the methodology can be assessed by cross-validation to give an insight into how the model will generalize to an independent dataset. In a following paper we will include a cross-validation procedure and a simulation study to assess the predictive accuracy and evaluate the performances of the algorithm.

In the next section, an application on real data is presented.

### 3 Application on real data

In this section a real data application of KM-LR is presented. The data set is named *Wine Data* [5]. It is the result of a chemical analysis of wines grown in an Italian region, derived from three different cultivars.

The 13 constituents were measured on 178 types of wine from the three cultivars: 59, 71 and 48 instances are in class one, two and three, respectively. The 13 chemical continuous attributes of the wine data set are: 1. *Alcohol* (Alc), 2. *Malic acid* (Mal), 3. *Ash* (Ash), 4. *Alkalinity of ash* (AAsh), 5. *Magnesium* (Mg), 6. *Total phenols* (Phe), 7. *Flavonoids* (Fla), 8. *Nonflavanoid phenols* (NPhe), 9. *Proanthocyanins* (ProA), 10. *Color intensity* (Col), 11. *Hue* (Hue), 12. *OD280-OD315 of diluted wines* (ROD), and 13. *Proline* (Pro).

In the analysis, we have tried to select the optimal number of clusters without considering the *a priori* information that  $K = 3$ , and using the KM-LR algorithm, i.e., through the maximization of the increase of the  $\chi^2$ -test computed on the partitions obtained by KM and LR. For comparison purposes, two other approaches have been used. The procedure has been randomly repeated 50 times from 2 to 10 clusters using a single random start. In Table 1, the results obtained by KM-LR (first column), the sequential application of KM followed by the *Gap-method* proposed by Tibshirani [12] (second column), and the sequential application of KM followed by Calinski and Harabasz's [2] criterion (third column) have been reported.

**Table 1** Optimal  $K$  selection from 2 to 10 clusters on the 50 random repeat using a single random start

K	KM-LR		KM → Gap-method		KM → Calinski-Harabasz	
	Count	Percent	Count	Percent	Count	Percent
2	0	0.00	0	0.00	0	0.00
3	36	72.00	5	10.00	22	44.00
4	10	20.00	0	0.00	5	10.00
5	2	4.00	0	0.00	3	6.00
6	2	4.00	0	0.00	3	6.00
7	0	0.00	2	4.00	0	0.00
8	0	0.00	1	2.00	0	0.00
9	0	0.00	15	30.00	6	12.00
10	0	0.00	27	54.00	11	22.00
Total	50	100.00	50	100.00	50	100.00

The best performance has been obtained by the KM-LR approach, where the optimal number of clusters has been captured 36 times out of 50 (72%) runs. In contrast, the KM-*Gap-method* obtained the worst performance, since the optimal number of clusters was captured only 5 times (10%). Thus, the KM-LR approach seems to reduce the effect of the local minima problem of the KM algorithm, and this is even more relevant in the case no modification of the KM partition as proposed by the KM → *Gap-method* and KM → *Calinski-Harabasz* method.

In Table 2 we show the estimation results of LR applied to the group labels identified through the KM model as a response variable and include only variables with significant coefficients as predictors.

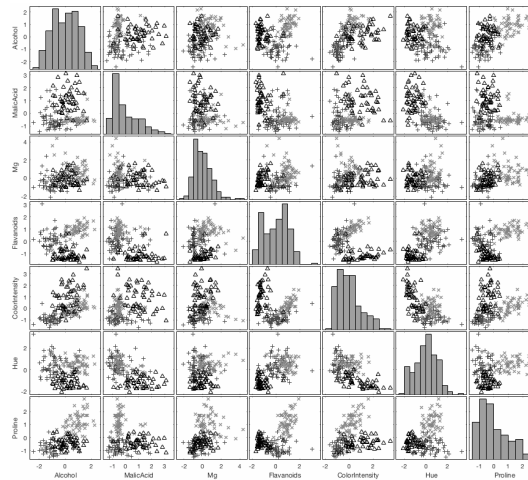
**Table 2** Estimation results obtained by logistic regression applied to the KM partition including only predictors with a 5% significant coefficient

	Estimate	SE	t-stat	p-value
Const.	2.0169	0.0296	68.2200	2.66E-122
Alc	-0.2306	0.0465	-4.9579	1.76E-06
Mal	-0.0865	0.0382	-2.2674	2.47E-02
Mg	-0.1264	0.0353	-3.5808	4.51E-04
Fla	-0.2012	0.0786	-2.5597	1.14E-02
Col	-0.0806	0.0516	-1.5634	1.20E-02
Hue	0.0970	0.0474	2.0492	4.20E-02
Pro	-0.3627	0.0498	-7.2806	1.31E-11

178 observations, 164 error degrees of freedom  
 Dispersion: 0.138, AICc=160.34, BIC=185.95  
 R-squared-adj.=0.8135  
 F-statistic: 93.70, p-value=5.19E-55

From Table 2 we can note that the model shows good performance and about 80% of the total deviance is explained (i.e.,  $R^2_{adj} = 0.81$ ). The variables *Ash*, *Alkalinity of Ash*, *Total phenols*, *Nonflavanoid phenols*, *Proanthocyanins*, and the *OD280-OD315 of diluted wines* have been excluded because these were not statistically significant at the 5% level. In Figure 1 the partitions identified by KM-LR (highlighted with different symbols) on the 7 included variables have been represented.

**Fig. 1** The 3 clusters identified by KM-LR represented on the variables included in the model



The partition seems well represented on most pairs of variables, this because it is represented by the statistically most significant variables. Moreover, the partition found by the KM-LR approach better identifies the real data partition identified by the three different cultivars.

Table 3 shows (i) the confusion matrix between the real data partition and the KM partition (i.e., KM applied to the complete data) and (ii) the confusion matrix between the real data partition and the KM-LR partition.

**Table 3** Confusion matrix between: (i) real data partition and KM partition; (ii) real data partition and KM-LR partition

Real	K-means			Total	Real	K-means - LR			Total
	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>			C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	
C <sub>1</sub>	32	5	22	59	C <sub>1</sub>	51	3	5	59
C <sub>2</sub>	9	61	1	71	C <sub>2</sub>	3	66	2	71
C <sub>3</sub>	2	27	19	48	C <sub>3</sub>	0	12	36	48
Total	43	93	42	178	Total	54	81	43	178

The misclassification rate and the adjusted Rand index (ARI) [11] applied on the left table (i.e., the real partition versus the KM partition) are equal to 0.3708 and 0.2977, respectively; these same indices applied to the right table (i.e., the real partition versus the KM-LR partition) are equal to 0.1818 and 0.5465, respectively. We recall that ARI has a value between 0 and 1, with 0 indicating that the two data clusterings do not agree on any pair of points and 1 indicating that the data clusterings are identical.

Moreover, by applying LR to the real data partition we obtain the following confusion matrix between the real partition and the one fitted by LR (Table 4).

**Table 4** Confusion matrix between real data partition and LR partition

Real	Logistic regression			Total
	C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	
C <sub>1</sub>	15	44	0	59
C <sub>2</sub>	6	62	3	71
C <sub>3</sub>	2	38	8	48
Total	23	144	11	178

The performance of KM-LR is also better. In fact, the misclassification rate and ARI applied to Table 4 are equal to 0.5225 and 0.0247, respectively. In Table 5, the performances obtained by both LR applied to the real partition and KM-LR are shown. We note that the diagnostic indices obtained by KM-LR are much than those obtained by the LR application on the real data partition. Furthermore, in the application of LR on the real data partition, only the variable *Color intensity* has obtained a statistically significant coefficient and then, only this variable has been included in the model.

**Table 5** Comparison between LR and KM-LR

	LR	KM-LR
<i>F</i> -Statistic	14.5000	93.7000
<i>p</i> -value	0.0002	5.19E-55
<i>R</i> -squared-adj.	0.0710	0.8135
AICc	403.3673	160.3400
BIC	409.6623	185.9500

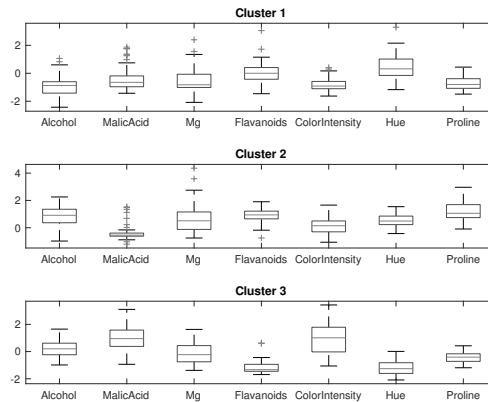
Finally, to obtain a quality measure of the clusters, a MANOVA model [10] on the real data partition and on that obtained by the KM and KM-LR models has been applied (Table 6).

**Table 6** MANOVA results obtained on the real data partition and on that obtained by *k*-means and *k*-means - logistic regression

	Wilk's Lambda	Chi-Squared approxima- tion	Degrees Freedom Chi-square	<i>p</i> -value	Partition
Const.	0.2052	267.6509	26	0.00E+00	Real
Group	0.7904	39.7581	12	7.89E-05	
Const.	0.2043	268.3793	26	0.00E+00	KM
Group	0.7609	44.3934	12	1.31E-05	
Const.	0.2303	248.1821	26	0.00E+00	KM-LR
Group	0.8063	36.3558	12	2.80E-06	

The null hypothesis is rejected in each of the three cases, i.e., the means of each group are not the same *j*-dimensional multivariate vector, and any difference observed in the sample is not due to random chance. However, we can note that the most significant value of  $\lambda$  is derived in the KM-LR partition. In Figure 2, the distributions of the three KM-LR clusters on the reduced set of variables are shown.

**Fig. 2** Boxplots of the three KM-LR cluster distributions represented on the variables included in the model



## 4 Concluding remarks

In the unsupervised classification approaches, the unknown number of clusters and the lack of assessment of the final partition are crucial issues that could negatively affect the reliability of the results. In this work we proposed an algorithm that combines KM and the LR modeling to evaluate the partition identified through KM, to assess the correct number of clusters, and to verify the selection of the most important variables. We did this by using well-known inferential tools that allowed us to statistically confirm the obtained results.

The application on real data shows that this methodology obtains better performance than the usual KM approach, reducing the effect of local minima. Moreover, KM-LR represents a useful tool to identify the variables that better contribute to defining the group structure in the data and removing the statistically non-significant variables from the model. In this way, we have a parsimonious set of variables that define the best partition of the data. Thus, the methodology seems promising. However, in a following work, we wish to better assess, using an extensive simulation study, the performance of the proposed methodology.

## References

1. Agresti, A., Kateri, M. Categorical data analysis. In International encyclopedia of statistical science, pp. 206-208 (2011)
2. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1), pp. 1-27.
3. Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference*.
4. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7), pp. 1895-1923.
5. Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
6. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Miscellaneous clustering methods*. *Cluster Analysis*, 5th Edition, pp. 215-255.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning*, pp. 485-585.
8. Hepner, G., Logan, T., Ritter, N., & Bryant, N. (1990). Artificial neural network classification using a minimal training set- Comparison to conventional supervised classification. *Photogrammetric Engineering and Remote Sensing*, 56(4), pp. 469-473.
9. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14) pp. 281-297.
10. Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, pp. 23-34.
11. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), pp. 846-850.
12. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), pp. 411-423.