Simultaneous unsupervised and supervised classification modeling for clustering, model selection and dimensionality reduction

Modellizzazione simultanea di metodi di classificazione non-supervisionata e supervisionata per classificare, validare il modello e ridurre la dimensione dei dati

Mario Fordellone and Maurizio Vichi

Abstract In the unsupervised classification field, the choice of the number of clusters and the lack of assessment and interpretability of the final partition by means of inferential tools denotes an important limitation 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, to identify the correct number of clusters and to select the variables that better contribute to define the groups structure in the data. An application on real data is presented in order to better clarify the utility of the proposed approach.

Abstract Nella classificazione non supervisionata, la scelta *a priori* del numero ottimale di gruppi da considerare e la mancanza di interpretazioni inferenziali, rappresenta un grosso limite per questi modelli. In questo lavoro, proponiamo la combinazione di modelli di classificazione non-supervisionata e supervisionata per identificare il numero ottimale di gruppi da considerare, selezionando le variabili che incidono in modo significativo sulla partizione trovata. E' prevista un'applicazione su dati reali.

Key words: Supervised Classification, Unsupervised Classification, Assessing Clustering, Model Selection, Dimensionality Reduction

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

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

1 Introduction

In the 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 heterogeneity structure of the data. Often, techniques based on separability and homogeneity criteria of the groups are used, giving *a priori* the number of groups [9].

Conversely, supervised classification is based on the idea to forecast 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, 6].

In this work, we propose a clustering algorithm based on the use of supervised classification modeling. In particular, the approach consists in the combination of K-Means (KM) and Logistic Regression (LR) modeling in order to find the correct number of clusters, select the most important variables and have an assessment on the partition identified through KM. An application on real data is finally proposed.

2 K-Means and Logistic Regression modeling into a clustering algorithm

In unsupervised classification modeling we are not interested in prediction, because we do not have an associated response variable y [5] like in a supervised classification model. The proposal of this paper consists in the combination of the unsupervised (i.e., K-Means (KM)) and supervised classification (i.e., Logistic Regression (LR)) approaches, where the latter, aiming to evaluate and to improve the former with adding data structure information. For simplify, we will call this approach K-Means - Logistic Regression (KM-LR). In particular, KM-LR is composed by the following principal steps:
Given the $n \times J$ data matrix $\mathbf{X}$, for $K = 2, \ldots, Kmax$, where $Kmax$ is the maximum number of clusters the researcher thinks the data might have, the algorithm works as follows:

1. let g_K be the unknown categorical membership variable which is estimated by using KM on the n-dimensional multivariate variables in $\mathbf{X}$ thus minimizing the objective function $\left\|\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\right\|^2$ [7];
2. g_K is used as response variable of the LR model with explanatory variables $\mathbf{x}$; LR is applied on g_K for estimating the probabilities for its $K-1$ response categories $\pi_k(\mathbf{x})$, and to estimate the probabilities for its *baseline* category $\pi_0(\mathbf{x})$, here fixed $k = 1$ [1];

3. if in second step some LR coefficient are not statistically significant, then we exclude the corresponding variables and we repeat from the step 1, otherwise $K = K + 1$.

Stopping rule: the algorithm continues until when the *Kmax* is reached. At the end, the optimal K is identified, together with a reduced set of statistically significant variables and a set of inferential tools to assess the quality of the partition.

In this way, through the analysis of the LR results (e.g., explained variance, parameters significance, residual variance, etc.) 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 well-explain the groups structure in the data. Moreover, through the LR coefficients analysis we can see which variables contribute most to identify the groups structure and to what extent they do it (then, analyzing statistical significance, estimates value, and sign of the coefficients).
In the next section, an application on real data is presented.

3 Application on real data

In this section a real data application of K-Means - Logistic Regression (KM-LR) is presented. The data set named *Wine Recognition Data*, is available at the UCI repository website (http://archive.ics.uci.edu/ml/). It is the result of the 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.

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 *chi-squared* test computed on the partitions obtained by K-Means (KM) and Logistic Regression (LR). For comparison purpose, other two approaches have been used. The procedure has been random repeated 50 times from 2 to 10 clusters. In Table 1 have been reported the results obtained by *chi-squared* (first column), *Gap-method* proposed by Tibshirani [10] (second column), and Calinski and Harabasz [3] criterion (third column).
Thence, the best performance has been obtained by KM-LR approach, where the optimal number of clusters has been captured 36 times on 50 (72%). Whereas, KM-*Gap-method* has been obtained worst performance, since the optimal number of clusters has been captured 5 times only (10%). Then, the KM-LR approach seems to reduce the effect of the local minima problem of the KM algorithm [2], which is more relevant in the case no modification of the KM partition is proposed as in the KM-*Gap-method*, and KM-*Calinski-Harabasz*.

In Table 2 the estimation results of LR applied on the groups labels identified through KM model as response variable and including only variables with significant coefficient as predictors are shown.

Table 1 Optimal K selection from 2 to 10 clusters on the 50 random starts

	Chi-squared		Gap-method		Calinski-Harabasz	
K	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

Table 2 Estimation results obtained by Logistic Regression applied on the K-Means partition including only predictors with 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
Ash	-0.1261	0.0438	-2.8778	4.54E-03
AAsh	0.1022	0.0444	2.3041	2.25E-02
Mg	-0.1264	0.0353	-3.5808	4.51E-04
Phe	0.0740	0.0617	1.1993	2.32E-01
Fla	-0.2012	0.0786	-2.5597	1.14E-02
NPhe	-0.0331	0.0397	-0.8326	4.06E-01
Pro	0.0885	0.0417	2.1243	3.51E-02
Col	-0.0806	0.0516	-1.5634	1.20E-01
Hue	0.0970	0.0474	2.0492	4.20E-02
ROD	-0.0832	0.0577	-1.4418	1.51E-01
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 a good performance, with about 80% of the total variance explained. In the model the variables *Ash*, *Alcalinty of Ash*, *Total phenols*, *Nonflavanoid phenols*, *Proanthocyanins*, *OD280-OD315 of diluted wines*, have been excluded because are not statistically significant at the 1% level.

Tables 3 show (i) the confusion matrix between real data partition and KM partition (i.e., KM applied on the complete data) and (ii) the confusion matrix between real data partition and KM-LR partition.

The misclassification Rate and the Adjusted Rand Index [8] applied on the left table (i.e., real partition versus KM partition) are equal to 0.3708 and 0.2977, respectively; whereas, the same indices applied on the right table (i.e., real partition versus KM-LR) are equal to 0.1818 and 0.5465, respectively.

Table 3 Confusion matrix between: (*i*) real data partition and *K*-Means partition; (*ii*) real data partition and *K*-Means - Logistic Regression partition

	K-Means			
Real	C_1	C_2	C_3	Total
C_1	32	5	22	59
C_2	9	61	1	71
C_3	2	27	19	48
Total	43	93	42	178

	K-Means - LR			
Real	C_1	C_2	C_3	Total
C_1	51	3	5	59
C_2	3	66	2	71
C_3	0	12	36	48
Total	54	81	43	178

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

Table 4 Confusion matrix between real data partition and Logistic Regression partition

	Logistic Regression			
Real	C_1	C_2	C_3	Total
C_1	15	44	0	59
C_2	6	62	3	71
C_3	2	38	8	48
Total	23	144	11	178

Also in this case the performance of KM-LR is better. In fact, the misclassification Rate and the Adjusted Rand Index applied on Table 4 are equal to 0.5225 and 0.0247, respectively. In Table 5 the performances obtained both LR applied on real partition and KM-LR are shown.

Table 5 Comparison between LR and KM-LR

	Logistic Regression	*K*-Means - LR
F-Statistic	14.5000	93.7000
p-value	0.0002	5.19E-55
R-Squared Adj.	0.0710	0.8135
AICc	403.3673	160.3400
BIC	409.6623	185.9500

We can note that the diagnostics indices obtained by KM-LR are very better with respect to those obtained by the LR application on the real data partition. Furthermore, note that 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.
In Figure 1 the distributions of the three KM-LR clusters on the reduced set of variables are shown.

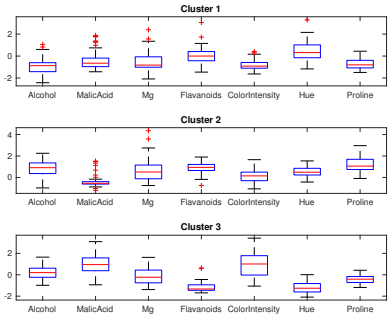

Fig. 1 Boxplots of the three KM-LR clusters distributions represented on the variables included in the model

4 Concluding remarks

In the unsupervised classification approaches, the choice of the 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 propose an algorithm that combines K-Means (KM) and the Logistic Regression (LR) modeling in order to have an evaluation of the partition identified through KM, assess the correct number of clusters (clustering) and verify the selection of the most important variables (model selection), removing in the model the non-significant variables (dimensionality reduction). In this way, we have a parsimonious set of variables that defines the best partition of data. Thus, the methodology seems promising, however, in a following work, we wish to better discover and assess, by an extensive simulation study, the performances of the proposed methodology.

References

1. Agresti, A., Kateri, M. Categorical data analysis. In International encyclopedia of statistical science, pp. 206-208 (2011)
2. Aloise, D., Deshpande, A., Hansen, P., Popat, P. NP-hardness of Euclidean sum-of-squares clustering. Machine learning, 75(2), pp. 245-248 (2009)
3. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), pp. 1-27.
4. Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7), pp. 1895-1923.
5. Filipovych, R., Resnick, S. M., & Davatzikos, C. (2011). Semi-supervised cluster analysis of imaging data. NeuroImage, 54(3), pp. 2185-2197.
6. 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.
7. 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.

8. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), pp. 846-850.
9. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining, 400(1), pp. 525-526.
10. 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.