ELSEVIER

Intelligent Data Analysis 3 (1999) 491–510

IDA
INTELLIGENT DATA ANALYSIS

www.elsevier.com/locate/ida

Evaluating a clustering solution: An application in the tourism market

Margarida G.M.S. Cardoso [a], Isabel H. Themido [a], Fernando Moura Pires [b,*]

[a] *CESUR, Inst. Superior Técnico, Univ. Técnica de Lisboa, Lisboa, Portugal*

[b] *Dep. Informática, Fac. Ciências e Tecnologia, Univ. Nova de Lisboa, Quinta da Torre, 2825-114 Monte de Caparica, Portugal*

Received 1 May 1999; received in revised form 9 August 1999; accepted 19 August 1999

Abstract

This paper discusses the evaluation of a clustering solution. Criteria based on the number of clusters and discrimination and classification processes are used to evaluate a clustering solution. The proposed approach is based on two paradigms: Statistics and Machine Learning. A multimethodological approach is advocated in the construction of models associating between properties and clusters, to provide a wider and richer set of analysis perspectives and a better knowledge discovery. Specifically, the construction of classification and discrimination logical models as a complement of quantitative statistical models is particularly useful when most of the available information is of a qualitative nature (nominal or ordinal variables). Both, the classification's global precision and the comprehension added by the discriminant model to the association between variables and clusters, are essential to evaluate a clustering solution. Depending on the dimension of the sample, descriptive analysis performed can be validated through the partition in two of the total sample – (one sub-sample for model build-up and another (holdout) for validation) – or by other procedures of cross-validation. The proposed evaluation approach is applied to a Marketing Tourism case study. The clustering solution is built upon a sample of more than 2500 Portuguese clients of *Pousadas de Portugal* Hotels. The database includes variables related to the evaluation of stay (per client) at the *Pousadas* and profiles of the surveyed clients on holidays, demographic and psychographic aspects. Measures of association, Chi-square tests, ANOVA, Discriminant Analysis, Logistic Regression, and Rule Induction (based on CN2 and C4.5 algorithms) are applied in evaluating the clustering solution built through a *K*-Means process.

Keywords: Clustering; Multivariate statistics; Machine learning; Marketing and tourism

1. Introduction

The main objective of this paper is to propose a framework, inspired in basic paradigms from Statistics and Machine Learning, for evaluating a clustering solution. A multimethodological approach is advocated. In particular, discriminant analysis and classification techniques, *latu sensus*, are considered in order to explain and predict/classify a nominal variable/cluster.

* Corresponding author.

E-mail addresses: margaridacardos@ip.pt (M.G.M.S. Cardoso), ithemido@civil18.civil.ist.utl.pt (I.H. Themido), fmp@di.fct.unl.pt (F.M. Pires)

1088-467X/99/$ - see front matter
PII: S1088-467X(99)00035-9

In the context of this work, *clustering* is referred to the partition of a set of entities in mutually exclusive sub-sets (discrete clusters), explicitly excluding wider interpretations of the term, such as fuzzy approaches and overlapping definitions. A general criterion is considered in the discriminant-classification evaluation of a clustering solution: the utility of the clustering solution is, in each practical application context, the dominating criterion of an evaluation (multi) methodology.

A case study illustrates the proposed approach: the clustering of a sample of Portuguese clients of *Pousadas de Portugal*, a network of more than 40 hotels scattered all over Portugal, is evaluated by the proposed methodology.

2. Criteria for the evaluation of a clustering solution

2.1. The various paradigms

A clustering solution can be evaluated in a dependence analysis perspective where the dependent variables are the clusters themselves. The input includes the cluster to which each entity belongs to together with other available attributes that characterise it. The objective is to perform an analysis to explain and predict the clusters.

There are several statistical methods available for this type of analysis, ranging from measures of association between variables, association tests, Automatic Interaction Detection (AID) [1], Classification and Regression Trees – CART [2], to Discriminant Analysis and Logistic Regression. Some of these methodologies fall in the domain of Statistical Inference, enabling the extension of the conclusions, given certain assumptions, to the population from which the sample of the clustered entities was drawn. Others, like the measures of association, belong to Descriptive Statistics.

Machine Learning broadens the nucleus of the statistical evaluation methods in the study of the relationships between the clusters and the characterisation of the entities themselves. This discipline proposes many Tree Classification algorithms, such as C4.5 [3], and specific algorithms for the induction of prepositional rules (for instance the CN2 [4,5]). These algorithms are particularly well fitted for the treatment of nominal variables and also have interpretable outputs.

Machine Learning methodologies are descriptive, as in Descriptive Statistics. Their conclusions are tested by empirical procedures of the cross-validation type. Given their nature, these procedures call for large databases, often with millions of records, enabling the distinction between robust patterns and mere coincidences. Giving particular attention to the computational complexity problem introduced by large databases, the new Data Mining techniques capitalise on the development of Machine Learning methodologies.

On evaluating a clustering solution, specific applications can make the best use of the different methodologies, sometimes in alternative, but preferably in tandem, resorting whenever adequate to inference methodologies complemented by quantitative or logic descriptive models.

In general, the conjugation of different methodologies, eventually based in different paradigms, sets the stage for dealing with rich and complex (real) problems [6]. When paradigms are not mutually exclusive, a multiparadigm methodology can be advocated. The adoption of a particular paradigm can be compared with the use of a specific instrument to observe the world (a telescope, for instance), each instrument being able to reveal information to which others are blind. As real problems are inevitably very complex and multidimensional and are usually tackled in stages, it is possible to use a different methodology in each phase.

2.2. Discrimination and classification

Discriminant analysis and classification analysis, *latu sensus*, can be described as the study of the association between the attributes describing a set of entities and clusters of those entities. The objective of this analysis is

- to make clear and explain the differences between clusters,
- to propose rules to classify individuals into clusters.

The base for this analysis may be the clustering variables themselves. In addition, alternative dimensions associated with the clusters may be included, reinforcing the discrimination and classification, and enhancing the insight into the clustering solution.

The discrimination analysis centres its attention in the models – functional, logical, graphical or otherwise – proposed in order to better discriminate among the clusters of a given system: providing evaluation criteria for the discrimination ability and giving special attention to matters related with the capacity to interpret the models.

Classification can be seen as the conclusion of the discrimination process: the entities subject to clustering are submitted to the rules proposed by the discriminant process (quantitative or otherwise) and the resulting classification is compared with the assignement associated with the clustering solution.

2.2.1. Statistical methodologies

Since clusters $C = \{c_1 \dots c_{|C|}\}$ can be considered as levels of a nominal variable, several statistical measures, and association processes and models used in relating nominal variables with other variables, can be considered in evaluating a clustering solution.

Measuring the degree of association between variables and clusters is a first step in the analysis. Different measures are proposed for variables of different types. For instance, the degree of association between two nominal variables A and C ($A = a_1, \dots, a_{|A|}; C = c_1 \dots c_{|C|}$) defined by V Cramer Statistic [10] is given by

$$V = \frac{\chi^2/n}{\sqrt{\min(|A|-1, |C|-1)}},$$

where n is the sample dimension, and

$$\chi^2 = \sum_{j=1}^{|A|}\sum_{j=1}^{|C|} \frac{\left(n_{a_i c_j} - (n_{a_i} n_{c_j}/n)\right)^2}{(n_{a_i} n_{c_j}/n)},$$

where $n_{a_i c_j}$ are the observed cross frequencies corresponding to a_i and c_j, n_{a_i} are the observed frequencies of a_i and n_{c_j} are the observed frequencies of c_j. V returns the value zero when there is no association between A and C and one if the association is perfect.

The Uncertainty Coefficient, U [10], adds to this type of measure of association, providing information about the importance of a certain variable (C) to explain variable A. U is based on a measure of entropy, $H(\cdot)$ and is defined by

$$U(A|C) = \frac{H(A) - H(A|C)}{H(C)},$$

where $H(A)$ is the entropy of the random variable A and $H(A|C)$ is the conditional entropy. Formally

$$H(A) = -\sum_{i=1}^{|A|} \frac{n_{a_i}}{n} \log_2 \frac{n_{a_i}}{n},$$

$$H(A|C) = \sum_{j=1}^{|C|} \frac{n_{c_j}}{n} H(A|C = c_j) = -\sum_{j=1}^{|C|} \frac{n_{c_j}}{n} \sum_{i=1}^{|A|} \frac{n_{a_i}}{n_{c_j}} \log_2 \frac{n_{a_i}}{n_{c_j}}.$$

The entropy assumes zero value if the observations are all concentrated in one of the possible levels of the random variable A and is maximum when the observations are equally distributed. The Uncertainty Coefficient, U, assumes a zero value if there is no association and the value one when the degree of association is maximal.

These concepts are also defined in Information Theory (see for instance [7]) where the concept of Mutual Information, $I(A;C)$ is defined as a measure of how a random variable can explain another random variable. This measure must verify the following relation:

$$I(A;C) = H(A) - H(A|C) = H(C) - H(C|A).$$

The Mutual Information, $I(A;C)$ is zero if there is no association among the random variables A and C, and assume the value $\min(H(A), H(C))$ when the degree of association is maximal, i.e., $H(A|C)$ or $H(C|A)$ are equal to zero.

The degree of association between two variables can be looked at from an inference point of view: the level of significance for the association of two nominal variables, for example, can be inferred from a sample using the Pearson chi-square of independence test.

For a quantitative variable, the ANOVA test measures the significance of the difference of the variable means for the various clusters. In this instance, the normality of the variable is assumed although the test is considered robust. The extension of this test for a vector of averages associated with several output variables of interest (MANOVA) also assumes the multinormality of the set of variables.

Both the measures of association and significance tests for association can be integrated in a classification tree. The CART proposal [2] is just an example of such an inclusion of statistical measures in a classification algorithm. In general, classification trees vertices represent subsets of entities and branches represent the ramification criteria based on variables. Their construction mirrors an association model – discrimination and classification – between the variables that characterise a sample and the clusters to which the entities of that sample belong.

The output of Discriminant Analysis, a traditional statistical technique for modelling the association between a set of variables and clusters, are linear functions of the variables that are able to discriminate between clusters. Assuming a multinormal distribution for the variables and significant differences of vector means between clusters (according to MANOVA test) it is possible to infer from a sample a model for the population. The criterion that guides the construction of the discriminant functions is the maximisation of the ratio between–within cluster variation.

Finally this synopsis of discriminant and classification statistical techniques would not be complete without Logistic Regression, a technique that considers as dependent variable, in a

linear model, the logarithm of the odds (ratio of the probabilities) of two clusters. In this analysis it is possible to make inferences without the assumption of the multinormality of explanatory variables, and to include qualitative variables in the set of explanatory variables. The criterion that guides parameter estimation in the construction of the model is the maximisation of the corresponding likelihood function.

2.2.2. Machine learning methodologies

Generally speaking, the evaluation of a clustering solution can also be seen as a supervised learning process, the input being the solution of an unsupervised learning process i.e., the clusters.

In Machine Learning there are several methodologies that try to extract knowledge from experiences (particularly on a classification/supervised learning perspective) and induce that knowledge to a wider universe from which the experiments are drawn. Some other different approaches, particularly those for which the knowledge representation induced is symbolic, are Decision Trees [2,3] and Propositional Rules [3,4]. Other techniques like Neural Networks or Naive Bayes [8,14] are available where the corresponding knowledge representation induced is sub-symbolic. In what concerns the choice of a particular approach we have to consider the problem in hand.

Classification trees, for instance, provide a hierarchical process and model of classification of entities belonging to clusters. They also provide a discriminant model of the logical type based on the attributes characterising the entities. The root node in the tree illustrates the group of all entities belonging to a mixture of clusters. Ramification illustrates the partitioning of entities in the nodes and is related to criterions that measure associations between the partitions associated with attribute levels and the clustering. On constructing a classification tree one is progressively revealing the cluster structure of entities.

Classification Trees' construction is generally, a nonbacktracking and greedy optimisation algorithm. However, in an attempt to improve the resulting solutions and overcome overfitting, several pre-pruning procedures (accomplished before tree completion) and post-pruning procedures can be used. The pruning techniques must take into account a trade-off between pruning at high levels of the tree (which can block future advantageous developments of the search) and pruning at low levels of the tree (which can originate complex results). Effects of pruning have to be considered in prediction, since in a terminal node a mixture of clusters (hopefully tending to degenerate in a particular cluster) will be present.

An alternative approach, the Induction of Propositional Rules, provides logic models for discrimination and classification of a clustering solution. Its output – a group of rules that can be represented by "*if* condition *then* cluster" – can be associated with classification trees' construction. In fact, the path from the root to a leave of a tree provides a conjunction of attributes' values, which can be considered in the "if condition" of a rule and the distribution of the clusters in the leaves provides insights for the classification. Alternatively, the search for rules can be conducted in the rules' space itself, considering successive specialisation of rules (/concretion of values for the attributes in the "if condition") and their impacts on classification [4,5].

2.3. Clustering base and complementary dimensions

Clustering is performed over a set of available information that measures relevant attributes on the entities to be clustered. These attributes constitute the clustering base. After obtaining a

clustering solution these same variables may be used in a different perspective, that of discrimination and classification. This analysis may bring some additional knowledge about the relative importance of the variables included in the base to differentiate between clusters. Additionally, precision of classification (the percentage of entities correctly classified) that results from the analysis is an indicator of the consistency of the clustering solution.

In alternative or in addition a clustering solution may be analysed in the same discrimination and classification perspective, but this time using external data. At this stage attributes that may add useful information about the profile of each cluster are included.

2.4. The number of clusters

The number of clusters may be set a priori or may be an outcome of the clustering process itself. In the first instance the number of clusters is based on external criteria (judgmental). Sometimes the output of the clustering process provides only a guide for the analyst's decision concerning the number of clusters, like in traditional hierarchical clustering where the dendogram plays a reference role. In general, the *best* number of clusters provided by a clustering process is obtained by comparing measures of model fit for as alternative numbers of clusters. For instance, in Mixture Models [11] there are several measures proposed to evaluate the number of clusters, most of them based on the likelihood function associated with the mixture model. In spite of the impossibility to assume a χ^2 distribution for the likelihood ratio function to infer conclusions these measures, deriving from information theory [11], are often used. The Akaike Criteria (AIC),

$$\mathrm{AIC} = -2\ln L + 2p$$

is an example where L is the likelihood function corresponding to the proposed mixture model (given the data) and p is the number of parameters estimated by the model. In general these criteria propose the maximisation of the likelihood function considering a penalty for models that estimate too many parameters.

Finally it should be noted that Discriminant and Classification analysis performed in the evaluation of a clustering solution, although indirectly, may encompass the number of clusters itself.

2.5. Clustering process and clustering solution evaluation

All clustering processes consider some kind of internal evaluation criterion that provides a concept of inter-clusters homogeneity and between cluster heterogeneity.

In this respect we can distinguish between two kinds of clustering processes and models that can be viewed as relevant in relation to the evaluation of clustering solutions. Clustering processes and models based on strictly *interdependency methods* provide a "common" internal evaluation of solutions basically related to between-inter cluster heterogeneity. In contrast *criterion based* methodologies for clustering (such as Mixture regression models, for example), may consider evaluation internally, not only seeking the between-inter clusters heterogeneity relation but also seeking for the adjustment of an intra-clusters dependency model. These latter methodologies evaluate clustering in a different and more complete perspective that can include covariables in the process of clustering itself. Hence, clustering solution evaluation, as proposed in this paper, is not so important for criterion based clustering methodologies.

2.6. Utility concepts

The main question in evaluating a clustering solution can be translated to a question about utility. Indeed, in evaluating a clustering solution, we are interested in finding several contributions for a concept of utility that can be considered general, area specific or application specific.

Generally speaking one can seek the occurrence of significant differences between clusters in relevant variables. The relevance of the variables can then be questioned in the specific domain and application context of the clustering solution.

Additionally Discriminant and Classification Models and Processes (again considering relevant variables) can highlight the utility of the clustering solution.

Finally utility is evaluated by judgement. For example, in a Management environment, the utility of a clustering solution can be judged, considering also the above-referred complementary analysis, by the help it provides to support decision making.

2.7. Proposed approach

Taking into account the above considerations, an approach for the evaluation of a clustering solution is proposed (see Fig. 1).

The analysis should be rooted in the context of each particular clustering application. It is in this specific environment that the a priori relevance of the clustering base and other external characteristics for evaluation should be considered.

In evaluating a clustering solution we have an input that includes clusters of entities and entities' characteristics. Attributes can be analysed in the light of their associations with the clustering solution. In this analysis the degree and significance of the association, as well as models and/or processes of association should be considered in order to add comprehension and consistency to the proposed solution.

A first step in the evaluation of a clustering solution should be to measure the association between considered variables and clusters. Then, the variables not significantly related with clusters should be discarded. Among significant relations observed, the eventuality of spurious relations occurring should be considered and perhaps more variables should be discarded.

Considering the results for the significance of associations we can prevent undesirable mix of variables in the output of our analysis by performing partial evaluation analysis considering, separately, groups of meaningful variables in the context of the particular application.

Discriminant and Classification Analysis *latus sensus* will provide the main output of the clustering solution evaluation process. Models of association of varying complexity between

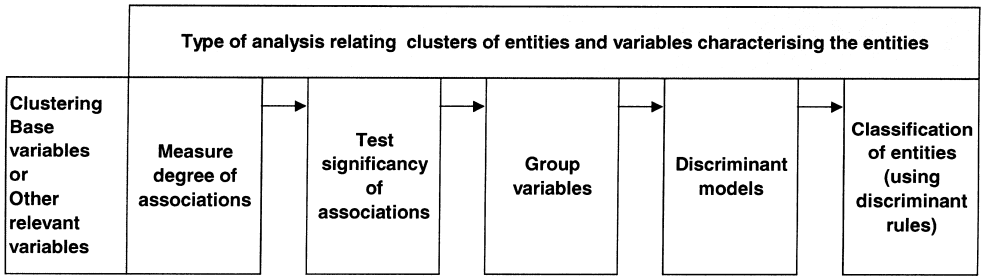

Fig. 1. Evaluation of a clustering solution.

attributes and clusters can be constructed. The results will emphasise the differences between clusters adding comprehension to them (discrimination) and will provide ways of predicting/classifying.

The choice of a particular discriminant and classification methodologies is guided by considerations about the nature of the variables considered (quantitative versus qualitative). The adjustment of the assumptions of the different methods to the data (for instance normality), the preference for a quantitative, logic or graphical model, or the size of the databases, will also guide the choice of methodologies.

In what concerns classification, the clustering base will most probably provide the best precision, depending on the discrimination–classification methodology and its relationship with the clustering process itself. Regarding discrimination, complementary dimensions will offer a new perspective and understanding of the clustering solution. This analysis will provide the means to prove and/or reinforce the consistency and interpretability and, finally, the utility of the clustering solution evaluated.

An integration of methodologies and techniques based mainly on the Statistical and Machine Learning Paradigms is proposed. This approach is based in the possibility that diverse and interesting methodologies can be combined to provide better outputs. This combination is performed in steps or in complementary approaches. An example is the combination of traditional Discriminant Analysis performed over quantitative variables (considering some assumptions) with Propositional Rule Induction Processes which is specially appropriate for the analysis of qualitative variables.

In the next part of this paper this approach is exemplified in a Tourism Market case study: the evaluation of a clustering solution for the Portuguese clients of *Pousadas de Portugal* Hotels Network.

3. A tourism market application

3.1. The clustering solution

3.1.1. Data base

The data used originated from the answers to a questionnaire directed to the Portuguese clients of *Pousadas de Portugal*. The questionnaire was specifically constructed to gather information for this market clustering study.

The questionnaires were distributed in all the the *Pousadas* between November 1996 and October 1997. The distribution was organised following a quarterly plan, each quarter including an equivalent group of heterogeneous *Pousadas* (Hotels).

The questionnaire included questions about the *Pousadas* (from motivation of stay and choice of the *Pousada*, to characterisation and evaluation of the stay and of the surrounding region), as well as questions about the holidays and the demographic and psychographic profile of the clients. The total number of questions was 49, translated, after codification and construction of new variables, into roughly 200 basic variables.

The sample collected under this project includes over 2500 Portuguese clients (a reply rate of 74%).

The respondent's average age is 45 years, with a standard deviation of 14. The respondents are mostly men (76%), married (87%). *The Empty Nest* category (over 35, married, no kids at home)

contains most (41%) of the clients. Most (62%) of the respondents have a higher education degree and high to very high income according to Portuguese standards (monthly net income of 1000–2500 Euro – 37% and 2500–5000 Euro – 33%).

3.1.2. Modelling and validation samples

In order to allow a correct evaluation of the clustering solution, the sample (2.544 clients) was previously divided into two subgroups, designated Model sample (Training set) and Validation sample (Test Set), respectively. The Model sample includes 1.647 clients (65% of the total sample) and the Validation sample includes the remaining cases (897 or 35% of the sampled clients).

The sub-samples extraction was done through the generation of random numbers associated with the sample individuals so that the proportional representation of the different *Pousadas* was respected and the Model sample was about two thirds of the total sample.

The clustering solution is simultaneously built in the Model and Validation samples. A discriminant and classification (*latu sensu*) analysis builds its models over the Model sample and evaluates its consistency and precision over the Validation sample cases.

3.1.3. Clustering

The clustering solution under analysis was built using a methodology already tested for a subsample of roughly 10% of the present sample [9], integrating clustering a priori and a *K*-Means procedure.

The clustering process defined three clusters for the Portuguese clients of *Pousadas*: *First time users*, *Regular users and Heavy users*, representing, respectively, 18%, 60% and 22% of the Model sample (for the validation sample 16%, 62% and 22%).

The cluster of *Pousadas*' *First time users was a priori* selected from the sample. The remaining respondents were then clustered, the clustering being based on four variables expressing the frequency and type of *Pousadas* (CH, CSUP, C and B types, roughly corresponding to a decreasing average price) [1] where the clients had already stayed overnight. The base variables for clustering were

- number of CH type *Pousadas*,
- number of CSUP type *Pousadas*,
- number of C type *Pousadas*,
- number of B type *Pousadas*,

where the client had already stayed overnight.

For clustering a *K*-Means procedure was used. Two clusters designated as *Heavy users* and *Regular users* were constituted. The first group of clients has already been lodged at an average number of 16 *Pousadas* and the latter at 5 *Pousadas*, according to values of the Model Sample (in the Validation Sample the *Heavy users* average increases to 17). The average number of each type of *Pousadas* where clients of each cluster client had already stayed overnight is shown in Fig. 2 (when data were collected *Pousadas* CH, CSUP, C and B were in number, 13, 4, 14 and 11, respectively).

[1] The classification of *Pousadas de Portugal* changed some months ago and now considers only two type of *Pousadas*: Regional and Historical. However, at time of data collection the mentioned categories were adopted and can still be considered relevant considering their relation to price.

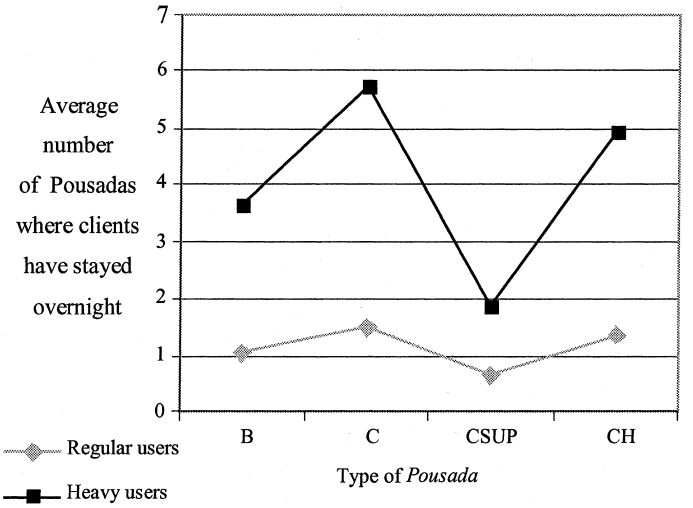

Fig. 2. Average number of each type of *Pousadas*, where clients of each cluster have stayed overnight (results from Model Sample).

3.2. Evaluation of the clustering solution

Evaluation of the clustering solution for the sample of Portuguese clients of *Pousadas de Portugal – First time users*, *Regular users* and *Heavy users* – was made in a discriminant and classification perspective. Fig. 3 resumes the main components of the evaluation procedure each particular analysis' component being described in the next parts of this presentation.

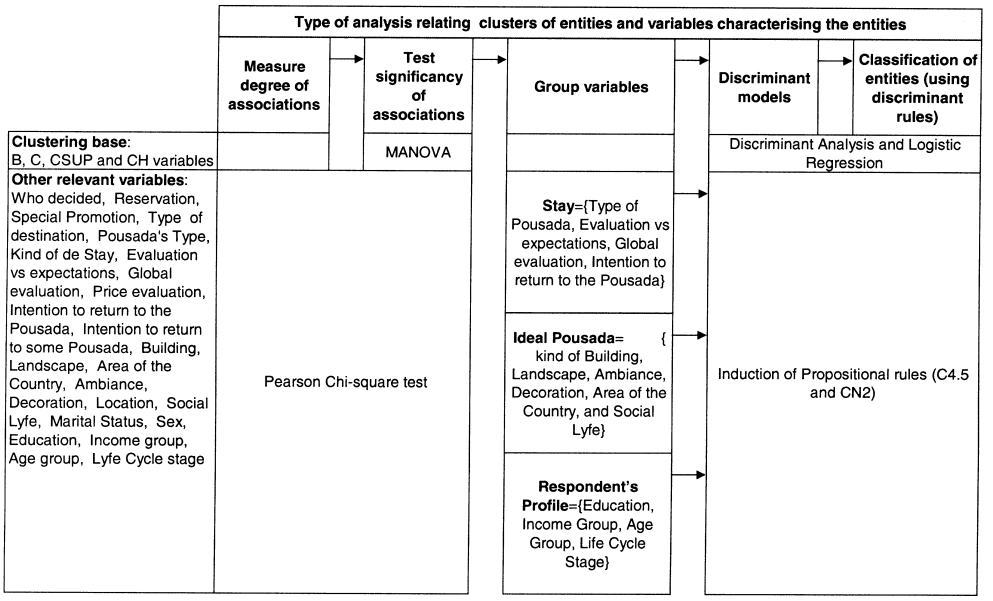

Fig. 3. Evaluation of clustering of Portuguese clients of *Pousadas de Portugal*.

The number of clusters was set a priori. This option was justified by the interest in separating, the first time users from the regular and differentiated users of *Pousadas*. The evaluation of the clustering solution subsequently performed indirectly validates the chosen number of clusters.

3.2.1. Analysis of association between clusters and clustering base

A Discriminant Analysis was conducted using the clustering base (B, C, CSUP and CH variables), after testing for significant differences between the clusters. In spite of the variables not being multinormally distributed a MANOVA test was conducted, considering the robustness of the *F*-test (particularly sensitive to outliers but less sensitive to deviations from symmetry as the current case).

The results from Discriminant Analysis reinforce the clustering solution proposed by *K*-Means, providing a precision for global classification of 95.5% in the Model sample (and 97.8% in the Validation Sample) that refers to the distinction between *Heavy* and *Regular users*.

The Huberty Index measures the degree of correction in classification, providing comparison with precision of classification according to majority (classifying all cases in the larger cluster). Its value for Model and Validation samples is 82.6% and 91.5%, respectively. Formally

$$\text{Huberty Index} = \frac{(\text{Pr} - \text{Pm})}{(1 - \text{Pm})},$$

where Pr is the percentage of correctly classified cases according to the analysis and Pm is the percentage of correctly classified cases according to the majority rule.

Such a result could have been anticipated since the objective of the *K*-Means Procedure is to maximise the variation between groups and to minimise the variation within groups and a similar objective is considered in the construction of linear discriminant functions: the linear combinations of the clustering base variables that maximise the ratio between–within cluster variation. The remaining results from Discriminant Analysis are presented in Table 1.

Logistic regression conducted over the logit function of the *Regular users* cluster's probability and *Heavy users* cluster's probability was performed as an alternative to Discriminant Analysis [12,13]. The former does not constrain base variables to multinormality in order to provide inference conclusions which is an advantage (although, as already mentioned, the discriminat analysis *F*-test can be considered robust).

The classification results provided by Logistic Regression show a global classification precision of 99.8% (Model sample).

The conclusions in what concerns the relative importance of variables for discrimination (similar from both the analysis) can be seen in Tables 1 and 2 revealing that consumption levels of CH and C types of *Pousadas* discriminate the most between *Heavy* and *Regular users*.

Table 1
Discriminant analysis results: *Regular users* vs. *Heavy users*

Discriminant function means		Coefficients (standardised) on discriminant function/loadings			
Heavy users	Regular users	B	C	CSUP	CH
−2.38	0.83	−0.29/−0.55	−0.55/−0.79	−0.19/−0.45	−0.5/−0.65

Table 2
Logistic regression results: *Regular users* vs. *Heavy users*

Logit equation coefficients (standardised)					$-2\ln L$			χ^2
Constant	B	C	CSUP	CH	Model with no variable	Model with variables	Difference	
−4.73	1.91	3.79	0.56	2.86	15.07	140.8	1366.2	0

3.2.2. Analysis of association between clusters and other variables of interest

The database contains information about the respondent's profile, the evaluation of stays and some variables revealing respondents' preferences, in particular those related with the concept of *Ideal Pousada*. All these variables, except age, are qualitative in nature. This fact invalidates performance of quantitative analysis like Analysis of Variance and Discriminant Analysis. The proposed methodology to deal with these variables begins with a chi-square test to evaluate the strength of association between clusters and variables of interest and then performs Rule Induction Procedures to discriminate and classify on the base of attributes significantly associated with clusters.

The C4.5 [3] and CN2 [4] algorithms were used to learn propositional rules over the application. The global precision of classification of the logic model (group of rules) provided by these algorithms, as well as the relevance and precision of each individual rule, was evaluated. The evaluation was performed not only with the Model sample (which is used to construct the logical model) but also with the Validation sample (reserved to the test of the model).

Rule induction was conducted over the complete clustering solution and considering pairs of clusters to provide a better comprehension of the facts discriminating the clusters.

3.2.2.1. Chi-square tests and cross-tab analysis. Chi-square tests were conducted to test associations between clusters and some variables of interest. Several significant associations were found (see Table 3). Note (in the Table 3) that some of the levels of the attributes were aggregated – worst/much worst in Evaluation vs Expectations, for example – in order to overcome situations where results of expectations are lower than 5 that restrict the application of the chi-square test. After, only significant associations with clusters were considered for further analysis, and the corresponding cross-tabs were analysed providing association's explanations (see Section 3.2.3).

3.2.2.2. Grouping variables. A selection of variables to support rule induction was made, choosing among those variables significantly associated with the clusters. Additionally, variables were grouped to improve the interpretability of the rules' output.

- A first group called *Stay* includes Type of *Pousada* where the respondent was staying, Evaluation of stay comparing with expectations, Global evaluation of quality of stay and Intention to return to the *Pousada*.
- A second group called Ideal *Pousada* includes the kind of Building, Landscape, Ambience, Decoration, Geographic location and Social life ideally associated with a *Pousada*.
- A third group includes characterisation of the respondent's *Profile*: Education, Income Group, Age Group and Life Cycle Stage.

Note: The levels of these variables considered for the analyse can be found in Table 3.

Table 3
Results of χ^2 tests

Group of variables	Variables	Levels of variables	$P(\chi^2)$
Stay	Who decided	Respondent; other	0.000
	Reservation	*Pousada*; Central; Travel Agency	0.000
	Special Program	Yes; No	0.000
	Type of destination	One Only; One of several; Stop over	0.676
	Pousada's type	B; C; CSUP; CH	0.000
	Kind of stay	Leisure; Work	0.559
	Evaluation vs. expectations	Much better; Better; Worst/Much worst	0.000
	Global evaluation	Very good; Good; Reasonable/Weak	0.001
	Price evaluation	Very expensive; Expensive; Fair; Cheap/Very cheap	0.500
	Intention to return to the *Pousada*	I would like very much; I would like; I wouldn't like very much/ I wouldn't like	0.030
	Intention to return to some *Pousada*	Certainly return; Probably return; May be/Probably not/ Certainly not	0.000
Ideal *Pousada*	Building	Monument; Regional	0.010
	Landscape	Beach; Country; City; Mountain	0.000
	Country's zone	North; Centre; South	0.170
	Ambience	Refined; simple	0.000
	Decoration	Rural; Modern; Classic; Antique	0.000
	Location	Inland; Coastland	0.000
	Social life	Reserved; Moderate; Lively	0.000
Profile	Marital status	Single; Married; Separated; Widow	0.000
	Sex	Male; Female	0.000
	Education	Primary; Basic; Secondary; Polytechnic; Degree; Master/ Ph.D.	0.000
	Income group	<1000; 1000 to 2500; 2500 to 5000; >5000 (Euro)	0.000
	Age group	<25 yr; 25–35 yr; 35–45 yr; 45–55 yr; 55–65 yr; >65 yr	0.000
	Life cycle stage	<35 yr and single; <35 yr, married and no children; <35 yr, married and young children; <35 yr, married and older children; >35 yr, married and children; >35 yr, married and no children at home; >35 yr, and living alone; other	0.000

3.2.2.3. C4.5 rule induction. In the C4.5 algorithm, construction of rules is based on a Decision Tree. This tree is constructed according to a top-down procedure in which ramification is based on a measure of association, the *Information Gain Ratio*:

$$\text{Information Gain Ratio} = \frac{I(C;A)}{H(A)}.$$

In C4.5 the decision tree is pruned in accordance with a classification precision criterion and rules' construction is based on the paths between the root and the leaves, working around the "if part" of the rule i.e., eliminating some attributes in the path according to some measure of classification' precision.

Eliminating those that do not add to global classification precision finally chooses a group of rules. For classifying new cases, rules are ordered and a default classification is considered (cases

Table 4
C4.5 Rules

Cluster			Groups of variables	Discriminant characterization			
Heavy users	Regular users	First time users		Total no. of rules (heavy; regular; first time)	Rules	Rules' discrimination	
						Model sample (*; **)	Validation sample (*; **)
x	x	x	Profile	20 (4; 4; 12)	Between 35 and 45 yr ⇒ Regular user	(272; 69.2%)	(149; 65%)
					Between 25 and 35 yr and University degree ⇒ Regular user	(97; 65.5%)	(65; 81.3%)
			Stay	5 (0; 4; 1)	Pousada CSUP ⇒ Regular user	(147; 69%)	(68; 66.7%)
			Ideal *Pousada*	6 (0; 3; 3)	Mountain Landscape and Rural Decoration ⇒ Regular user	(155; 69.5%)	74; 58.3%)
					Monument, Inland location and Moderated Social life ⇒ Regular user	(243; 67.1%)	(127; 65.1%)
x	x		Profile	11 (2; 9;–)	More than 65 yr and University degree ⇒ Heavy user	(43; 58.9%)	(35; 71.4%)
					25–35 yr ⇒ Regular user	(218; 89%)	(118; 88%)
					35–45 yr ⇒ Regular user	(260; 82.3%)	(135; 77.6%)
			Stay	4 (0; 4;–)	Pousada CSUP ⇒ Regular user	(147; 84.5%)	(68; 76.4%)
					Pousada B ⇒ Regular user	(254; 80.1%)	(146; 81.6%)
			Ideal *Pousada*	3 (0; 3;–)	Mountain Landscape and Simple Ambiance ⇒ Regular user	(190; 84%)	(91; 79.8%)
	x	x	Profile	10 (–; 6; 4)	University degree and Income de 2500 to 5000 Euros ⇒ Regular user	(170; 84.2%)	(90; 87.4%)
			Stay	5 (–; 5; 0)	Expected Quality ⇒ Regular user	(390; 77.5%)	(212; 81.2%)
					Pousada C ⇒ Regular user	(285; 82.8%)	(47; 83.9%)
			Ideal *Pousada*	5 (–; 2; 3)	Reserved social life ⇒ Regular user	(257; 84.3%)	(142; 87.1%)
					Monument and Inland location ⇒ Regular user	(300: 82%)	(157; 82.6%)
x		x	Profile	12 (6;–; 6)	More than 65 yr ⇒ Heavy user	(66; 93%)	(46; 93.9%)
					Between 55 and 65 yr ⇒ Heavy user	(81; 89%)	(33; 86.8%)
			Stay	7 (4;–; 3)	*Pousada* B ⇒ First time user	(120; 69.4%)	(62; 65.3%)
					Pousada C ⇒ Heavy user	(111; 71.2%)	(59; 66.3%)
					Pousada CH ⇒ Heavy user	(120; 61.5%)	(59; 67%)
			Ideal *Pousada*	12 (7;–; 5)	Rural Decoration and Moderated social life ⇒ First time user	(69; 75.8%)	(40; 71.4%)
					Refined Ambiance and Reserved social life ⇒ Heavy user	(93; 80. 9%)	(50; 80.5%)

* No of correctly classified examples.
** % of covered examples correctly classified.

Table 5
Precision of classification: C4.5 and CN2 results

Algorithm	Segments			Groups of variables	Classification precision							
					Model sample				Validation sample			
	Heavy users	Regular users	First time users		Classification precision/Class (Heavy users, Regular users, First time users) (%)	Global precision of classification (%)	Default precision of classification (by majority) (%)	Huberty index (%)	Classification precision/ Class (Heavy users, Regular users, First Time users) (%)	Global precision of classification (%)	Default precision of classification (by majority) (%)	Huberty index (%)
CN2	x	x	x	Profile	(25.5; 94.1; 18.4)	65.6	60.4	13.1	(20.3; 87.8; 9.8)	60.0	62.2	−5.8
				Stay	(2.9; 98.1; 9.4)	61.6		3.0	(1.6; 97.6; 5.6)	62.0		−0.5
				Ideal Pousada	(5.6; 98.3; 11.7)	62.7		5.8	(1; 94.9; 3.5)	59.8		−6.3
	x	x		Profile	(26.4; 95.7; –)	77.8	74.1	14.3	(24; 89.7; –)	72.7	74.2	−5.8
				Stay	(2.9; 99.7; –)	74.7		2.3	1.6; 100; –)	74.6		1.6
				Ideal Pousada	(11.1; 97.9; –)	75.4		5.0	(3.6; 96.6; –)	72.5		−6.6
		x	x	Profile	(–; 97.7; 23.4)	80.3	76.6	15.8	(–; 95.6; 11.2)	78.2	79.4	−5.8
				Stay	(–; 98.1; 9.7)	77.4		3.4	(–; 97.8; 5.6)	78.8		−2.9
				Ideal Pousada	(–; 98.8; 13.7)	78.8		9.4	(–; 95.6; 3.5)	76.7		−13.1
	x		x	Profile	(81.8; –; 78.6)	80.3	53.3	57.8	(80.7; –; 74.1)	77.9	57.3	48.2
				Stay	(76.8; –; 62.9)	70.3		36.4	(68.2. –; 52.4)	61.5		9.8
				Ideal Pousada	(75.4; –; 67.9)	71.9		39.8	(76; –; 65.7)	71.6		33.5
C4.5	x	x	x	Profile	(22.3; 89.9; 22.1)	63.3	60.4	7.3	(15; 89.7; 15.4)	61.5	62.2	−1.9
				Stay	(1; 95.4; 12.7)	60.7		0.8	(0; 98.4; 7.7)	62.4		0.5
				Ideal Pousada	(25.2; 82.9; 9.7)	61.2		2.0	(0; 96.9; 5.6)	61.2		−2.6
	x	x		Profile	(28.4; 90.5; –)	75.2	74.1	4.2	(18.2; 92.8; –)	75.8	74.2	6.2
				Stay	(0; 100; –)	74.1		0.0	(0; 100; –)	74.2		0.0
				Ideal Pousada	(28.7; 84.3; –)	74.1		0.0	(0; 100; –)	74.5		1.2
		x	x	Profile	(–; 90.2; 27.1)	78.0	76.6	6.0	(–; 98.2; 7)	79.4	79.4	0.0
				Stay	(–; 92.6; 18.7)	76.6		0.0	(–; 100; 0)	79.4		0.0
				Ideal Pousada	(–; 98.3; 9.7)	77.5		3.8	(–; 96.9; 5.6)	78.1		−6.3
	x		x	Profile	73.9; –; 84.3)	79.5	53.3	56.1	(76.6; –; 79)	77.6	57.3	47.5
				Stay	(74.2; –; 57.9)	66.6		28.5	(66.7; –; 55.9)	62.1		11.2
				Ideal Pousada	(63.6; –; 73.9)	68.9		33.4	(57.3; –; 76.2)	65.4		19.0

not covered by any rule are allocated to the larger cluster). The final output – group of rules – is very easy to read and interpret.

In the current application C4.5 was run based on the different groups of variables referred in Section 3.2.2.2. The corresponding results are presented in Tables 4 and 5. On the first table some rules that discriminate between clusters are presented. Rules presentation is associated with the number of examples correctly classified by the rule and percentage of covered examples (satisfying the *if part* of the rule) correctly classified. The rules revealed new insights into the relationship between the clusters and the variables included in the analysis.

Table 5 resumes the classification results. These results show weak precision of classification, with values near (occasionally lower than) those obtained with majority classification. In an attempt to improve these results a sub sample was considered where the proportions of individuals in the clusters were levelled, the representatively of the groups of clients staying in the different *Pousadas* being preserved. However this procedure did not achieve better results.

Results from the analysis considering pairs of clusters are better than those for the complete clustering solution. In particular, when the analysis refers to comparison between *Heavy* and *First time users* in the Profile group of variables the Huberty Index is 56%, meaning that more than half of the total possible increase in classification precision (having majority classification as a reference) was achieved.

3.2.2.4. CN2 rule induction. The CN2 algorithm learns by a top-down and beam search in the rules' space: it searches for increasingly specific rules (more and more concretions of attributes' values in the *if part* of the rule) and memorises a group/beam of the best solutions considering a Laplace's Heuristic maximisation criterion:

$$\textit{Pr}\text{ecisision of Laplace} = \frac{N_r + 1}{N + C},$$

where *Nr* is the number of cases correctly by the rule, *N* the number of cases covered by the rule and *C* is the number of clusters.

In the current application CN2 was run based on the different groups of variables referred in Section 3.2.2.2. Its results provided better precision for classification results than C4.5, for the Model Sample: the Huberty Index is, in average, 5 points higher for CN2 than for C4.5 results. However, it should be taken in consideration that pruning as implemented in C4.5 is more drastic than pruning conducted in CN2, and, in consequence, the overfitting effects are more relevant in CN2 as suggested by results achieved for the Validation Sample: in average the Huberty Index is 2.4 points higher for C4.5 than for CN2.

In what concerns discriminating rules CN2 produces an excessive number of rules (again a consequence of light pruning in the rules space), a fact that strongly penalise the interpretability of the output. In fact, although the precision of rules (percentage of covered cases correctly classified) is high the number of cases correctly classified by a rule is very little. It should also be noted that CN2 does not provide results for individual rules' precision in the Validation sample (it only provides global classification results).

3.2.3. Global evaluation

In Discriminant Analysis and Logistic Regression the results highlight more clearly the differences between clusters. These differences, observed on the variables used for clustering, underline the role of consumption of C and CH types of *Pousadas* in discrimination.

In what concerns other variables of interest (not used in the clustering process), some conclusions about the associations with the clusters – *Heavy*, *Regular* and *First time users* of *Pousadas*– are presented below. Results from chi-square tests provide identification of significant association between several of these variables and the clusters. The corresponding analysis of cross-tables also provides some interpretation of those associations. Finally the induction of propositional rules, by means of C4.5 and CN2 algorithms, provides a more complex and richer perspective proposing a logical model for the associations.

The conclusions presented in Tables 6–8 highlight several interesting associations between clusters and attributes characterising the stay in the *Pousadas*, the profile of respondents and Ideal concepts of *Pousadas*. These associations reinforce the clustering solution that was made upon criteria based on type and frequency of use of products *Pousadas*.

Rule induction, in particular, provided new insights in the application. It was possible to induce some rules with a good precision, measured by a large percentage of correctly classified cases and a reasonable number of cases covered by the rule and correctly classified.

In a classification perspective good results were achieved by Discriminant and Logistic Regression Analysis that provided a global precision of classification surpassing 95%. However the

Table 6
Clustering solution and the respondent's profile

Significant associations	Average ages of *First time users, Regular users* and *Heavy users* are, respectively, 37, 44 and 53 yr, and, according to ANOVA's results these values are significantly different (considering a 0.01 significance level). Marital status, Sex, Age group, Life cycle stage, Education and Income group show significant associations with clusters (see Table 3)
Cross-tab analysis	Single individuals can be found more frequently among *First time users* (17%) than among the rest of the clients (4% and 7% for *Heavy users* and *Regular Users* respectively). The percentage of women respondents is larger for the First time users (33%) and smaller among *Heavy users* (18%) The most frequent Age group among *First time users* is 25–35 yr of age (43%) Secondary and Polytechnic education are the education levels more common among *First time users* (46%). 69% of *Heavy users* have University degrees *Heavy users* have higher incomes. For *Heavy users*, *Regular users* and *First time users* the percentages for income group above 2500 Euros, are 69%, 44% and 32%, respectively
Propositional rules	Older than 65 yr → *Heavy users*. Rule produced by C4.5 analysing *Heavy users* vs. First Time users, which has 93% precision and applies correctly to 66 individuals Age between 25 and 35 yr → Regular Client. Rule produced by C4.5 analysing *Heavy users* vs. *Regular users* which has 89% precision and applies correctly to 218 individuals More than 65 yr, married, and no children at home → *Heavy users*. Rule produced by CN2, analysing *Heavy users* vs. First Time users, which has 91% precision and applies correctly to 46 individuals Between 25 and 35 yr and Polytechnic Education → First time users. Rule produced by CN2, analysing *Heavy users* vs. *First time users* which has 94% precision and applies correctly to 30 individuals

Table 7
Clustering solution and the stay in the *Pousada*

Significant associations	Who chose the *Pousada*, the Type of Reservation, Special Promotion, Type of Destination, *Pousada's* Type, Evaluation vs. Expectations, Global Evaluation and Intention to return to some *Pousada* are variables that show significant associations with clusters (see Table 3)
Cross-tab analysis	*Heavy users* (93%) and *Regular users* (85%) choose the *Pousada* themselves while *First time users* rely more on others' choice (although 77% are still responsible for the choice) stay The Reservations' Central is more used by *Heavy users* (41% vs. 30% for the other clusters) and Travel Agencies are comparatively more used by *First time users* (15% vs. 10% and 3% for *Regular* and *First Time users*, respectively) *Heavy users* benefit more from Special Promotions (45%). For *Regular users* and *First time users* percentages are 21% and 12%, respectively *Heavy users* go more frequently to CH type *Pousadas* (38% vs. around 27% for the rest of the clients); *First time users* can be found more frequently in B type *Pousadas* (40% vs. 16% and 26% corresponding to *Heavy users* and Regular users, respectively) More *First time users* are surprised by quality of stay: 45% positively surprised. The majority of the other clients (around 65%) find in the *Pousada* what they expected *First time users* appreciate more the quality of their stay: 98% rank it in Very Good or Good, although this percentage is also very high for the rest of the clients (90%) *First time users* show higher rates of intention to return to the *Pousada* where they are staying: 37% would like very much to come back and around 28% of the remaining respondents declare similar intention *First time users* are not so shure of returning to a *Pousada* of Portugal as the other clusters: "I will certainly return" is an intention shared by 60% of First time users, 88% of *Regular users* and 96% of *Heavy user*
Propositional Rules	Stay in CH type Pousada → *Heavy users*. Rule produced by C4.5 analysing *Heavy users* vs *First time users* which has 62% precision and applies correctly to 120 individuals Quality of stay equals the expected → Regular Client. Rule produced by C4.5 analysing *Regular users* vs. *First time users* which has 78% precision and applies correctly to 390 individuals Stay in B type *Pousada* → First time users. Rule produced by C4.5 analysing *Heavy users* vs. *First time users* which has 69% precision and applies correctly to 120 individuals

classification results of rule induction procedures associated with the complete clustering solution, did not improve default classification precision.

As a conclusion the evaluation procedure for the clustering solution improved its comprehension and ability to support future Marketing decisions concerning the Portuguese clients of *Pousadas de Portugal*.

4. Conclusions and perspectives

This work main concern was the evaluation of a clustering solution. The proposed multimethodology associates Statistical and Machine Learning Paradigms. The evaluation was proposed in a (*latu sensu*) discriminant and classification analysis perspective, with the following objectives:

- to clarify and interpret the differences between the clusters;
- to construct classification rules of individuals into the clusters.

The proposed clustering evaluation approach was used in a Tourism Market Application.

Table 8
Clustering solution and the concept of Ideal *Pousada*

Significant associations	Building, Landscape, Ambience, Decoration, Location and Social Lyfe are components of the *Ideal Pousada* concept which show significant associations with clusters (see Table 3)
Cross-tab analysis	Although the majority of *First time users* prefer Monumental buildings their preference for the alternative, Regional Buildings, is relatively high: 35% vs. 30% and 24% corresponding to *Regular* and *Heavy users*, respectively The majority of clients (around 60%), prefer a *Pousada* on a Mountain Landscape. But in what concerns relative preferences, *Heavy users* show higher preference for *Pousadas* on a Country or city landscape and *First time users* show higher preference for *Pousadas* on a Beach Landscape (8% vs. 2% e 3% for *Heavy users* and *Regular users*, respectively) *Heavy users* show higher preference for refined ambience: 73% vs. 62% and 52% of *Regular users* and *First Time Users*, respectively *Heavy users* show higher preference for Classic or Antique Decoration and less preference for Rural Decoration: 35% of *Regular users* and 43% of *First time users* prefer a Rural Decoration and only 18% of *Heavy users* show identical preference *First time users* show a relatively higher preference for Coastland location: 28% vs. 19% and 12% of *Regular users* and *First time users*, respectively *Heavy users* show higher preference for Reserved Social Life (44%) while the other clusters prefer Moderate Social Life. However *First time users* show still some preference for Animated Social Life (12%)
Propositional rules	Preference for Rural Decoration and Moderated Social Life ⇒ First time users Rule produced by C4.5 analysing *Heavy users* vs. *First time users* which has precision 76% and applies correctly to 69 individuals

The use of tests for identifying significant associations between clusters and attributes characterising the clustered entities guided discriminant and classification analysis. The grouping of characteristics to be used in these analysis (rule induction, for example) empirically proved to be a strategy having a positive contribution to the quality and interpretability of the outputs.

Propositional rule induction was found particularly suitable for discriminating purposes, since most of the available variables were nominal and the intelligibility of its outputs/rules is appealing from a Marketing manager's point of view. Moreover it should be noted that rule induction provides a systematic search conducted over matrix data (attributes × clusters) that can learn useful concepts and goes far behind the analyst look over the traditional cross tables.

The results of CN2 (a particular rule induction procedure) in this application originated an overwhelming number of rules with a corresponding reduced number of cases that each rule applies to. This fact gave origin to some reflections about the trade-off between good discrimination provided by individual rules and the global precision of the classification provided by the entire group of rules. This trade-off should probably be considered in the internal evaluation process of the rule induction algorithms providing ways to give special relevance to interpretability in applications (like Marketing) for which this concern is specially important.

Finally, we advocate that a multimethodological approach to evaluate a clustering solution should consider not only inference but also descriptive analysis. The former type of analysis is able to extend conclusions to the population if some assumptions concerning the variables involved are verified. The latter one can rely on empirical cross-validation type procedures, to generalise its conclusions, based on big enough samples.

Acknowledgements

This research project was partially funded by *Pousadas de Portugal*, by the Portuguese Government through the *Ano Nacional do Turismo* (1996) and by Praxis XI programmes.

References

[1] G.V. Kass, An exploratory technique for investigating large quantities of categorial data, Applied Statistics 29 (2) 119–127.
[2] L. Brieman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth, 1984.
[3] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[4] P. Clark, R. Boswell, Rule Induction with CN2: Some Recent improvements, in: Proceedings of the Sixth European Working Session on Learning, Springer, Porto, Portugal, 1991, pp. 151–163.
[5] P. Clark, T. Niblett, The CN2 induction algorithm, Machine Learning 3 (1989) 261–283.
[6] J. Mingers, J. Brocklesby, Multhimethodology towards a framework for mixing methdologies omega, International Journal of Management Science 25 (5) (1997) 489–509.
[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[8] T. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.
[9] M.G.M.S. Cardoso, I.H. Themido, lientes portugueses das Pousadas de Portugal: um estudo de segmentação, Revista Portuguesa de Marketing 2 (6) (1998) 45–63.
[10] R.A. Cooper, A.J. Weeks, Data Models and Statistical Analysis, Philip Allan, Oxford, 1983.
[11] Richard P. Bagozzi, (Ed.), Advanced Methods of Marketing Research, Blackwell Business, 1994.
[12] S.J. Press, S. Wilson, Choosing between logistic regression and discriminant analysis, Journal of the American Statistical Association 73 (364) (Applications Section) 1978.
[13] B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, Journal of the American Statistical Association 70 (352) (Theory and Methods Section) 1975.
[14] P. Langley, Elements of Machine Learning, Morgan Kaufman, Los Altos, CA, 1996.