FORMAL DESCRIPTIVE STUDY FOR THE EXTRACTION AND COMPARISON OF TOURIST SPENDING PATTERNS

A. RABASA, N. MOLLÁ-CAMPELLO & A. PÉREZ-TORREGROSA
Operation Research Center, University Miguel Hernández of Elche, Spain.

ABSTRACT

This paper presents the design of an in-depth descriptive analysis of data collected from public surveys at tourist information points. It uses a dataset that compiles different information related to trips to the Valencian Community (Spain). The aim of this study is to describe the patterns (association rules) that a certain type of expense has, and how this could be used to improve the services offered to tourists.

There are different kinds of expenses to analyze: transport, accommodation, leisure as well as total daily expenses and total daily expenses per person. Those cases where expenses are especially high or low are considered as particularly important because of their strategic interest for the public administration of tourism.

The study starts with data preprocessing, followed by pattern extraction for the sub-samples with very high and very low expenses, and in some cases, zero expenses are not considered as outliers but as a particular group of individuals. After this, the study aims to extract the most important attributes (feature selection) to create a classification model and compare its efficiency with the models that compute the complete set of attributes.

To conclude, this paper presents the possible future predictive models that could lead to an improvement in planning for public tourist services in the Valencian Community (Spain).

Keywords: Feature Selection, Pattern Discovery, Predictive Tourism Analysis.

1 INTRODUCTION AND OBJECTIVES

As the authors have illustrated in a previous study [1], the Valencian Community (Spain) is one of the main tourist destinations for foreign visitors. It is also an area where the tourism offer is growing and diversifying enormously. In that paper, the authors examined tourist spending segmentation from a descriptive approach, by using clustering models that grouped tourist spending according to their travel motivation (leisure, business…), transport and accommodation. It provided very useful information for tourist managers in order to help them to design accurate tourism offers.

In the same application framework, this paper examines a predictive approach, trying to create qualitative models that can classify tourist spending according to their corresponding travel patterns, which are extracted from public surveys (EGATUR) [2].

There are several classification algorithms (Table 1) that can be characterized with different splitting criteria and admit different attribute type or different pruning strategy.

CART [3] is one of the most applied algorithms in the classification field (discrete response or target variables) and regression (continuous target variables). One important aspect of CART is that it can do attribute selection for splitting and selecting those variables that provide higher information gain (Gini Entropy). However, different attribute selection techniques have been applied successfully in the tourism sector, such as Multiple Criteria Decision Making for hotel sites [4], Principal Component Analysis [5] or Support Vector Machine on Tourism Recommendation Systems [6].

In this paper, we raise the following question: from a descriptive model (always required) for pattern search, is it possible to extract a set of attributes that could improve the model's accuracy or obtain the same accuracy as those that use all the variables?

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V13-N3-272-280

Table 1: Some decision tree models

	Splitting Criteria	Attribute type	Pruning Strategy
ID3	Information Gain	Only Categorical value	No implemented
CART	Gini Entropy	Categorical and Numeric value	Cost-Complexity
C4.5	Gain Ratio	Categorical and Numeric value	Error Based

These predictive models will be completely oriented towards establishing the extreme expense values (minimum and maximum), which are the most relevant aspects concerning public managers in the tourism sector.

2 DATA SET AND METHODOLOGY

2.1 Data Set

After data pre-processing and erasing the outliers, the global data set used for this study has a total number of 22742 records with 285 variables. From this global data set, several target variables are selected, including expenses on leisure, transport, accommodation or total expense, and a concrete data set is created for each expense using the minimum and maximum discretized ranges for each one. Thus, different files are created with different target variables (expenses on leisure, accommodation, transport, total), which are used to analyse the selected characteristics for each case.

Several subsets are created using different criteria to establish the maximum and minimum values for each type of expense (Table 2).

2.2 Methodology

The algorithms used for this study are CART and A priori, combined in order to improve the performance of the Classification And Regression Trees (CART) algorithm and to define the most important characteristics to establish extreme expenses (minimum and maximum expenses).

The generated subsets for expenses are trained and classified by the CART algorithm and obtain models with original accuracies, which will be compared with the accuracy of the new models trained by the CART algorithm with a previous attributes selection (using a priori algorithm). Thus, the methodology proposes to extract the 100 most reliable association rules from the minimum values and the 100 most reliable rules from the maximum values for each

Table 2: Summary of maximum and minimum levels of expenses

Expense	Criteria min	Criteria max	Values min	Values max	total minmax subset (% dataset)
Transport	0€–10€	> 100€	3877	2414	6291 (27.66%)
Accommodation	No expense	> 40€	11335	4816	16151 (71.02%)
Leisure	0€–33€	59€–80€	4460	4728	9188 (40.4%)
Total per day	0€–60€	> 200€	4117	4025	8142 (35.8%)

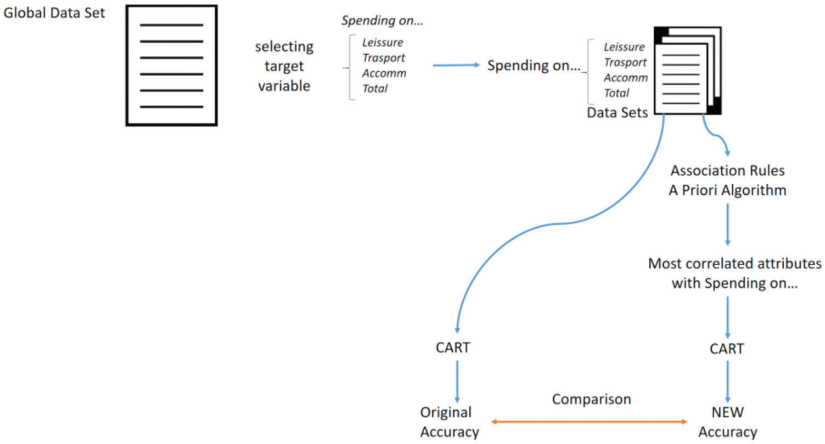

Figure 1: Summary of methodology

type of expense with the a priori algorithm. Those attributes that appear the greatest number of times in those rules are considered the most correlated variables with that type of expense and are selected to be trained by the CART algorithm in a new model, whose accuracy will be compared with the accuracy of the initial model.

3 COMPUTATIONAL EXPERIMENTS

The empirical results consist of performance estimates of the 2 classifiers across the data sets in terms of the four performance measures. Interested readers can find these raw results in Tables 2–10 and in the summary in Table 11.

3.1 Spending on leisure

Classification models that consider all attributes obtain an accuracy rate of 42.95%, a Type I error of 23.35%, a Type II error of 33.70% and root mean square error of 1.23. Table 3 shows the full confusion matrix.

The most frequent attributes in the association rules for leisure expenses (Fig. 2) are the person who paid the transport for the journey (GT2_4_2).

Table 3: Confusion Matrix for original leisure classification model.

	Very Low: 0–33	**Low: 34–43**	**Middle: 44–58**	**High: 59–80**	**Very High: >80**
Very Low: 0–33	840	306	200	82	33
Low: 34–43	76	176	68	4	6
Middle: 44–58	311	706	738	392	237
High: 59–80	55	101	190	458	265
Very High: >80	56	84	238	482	718

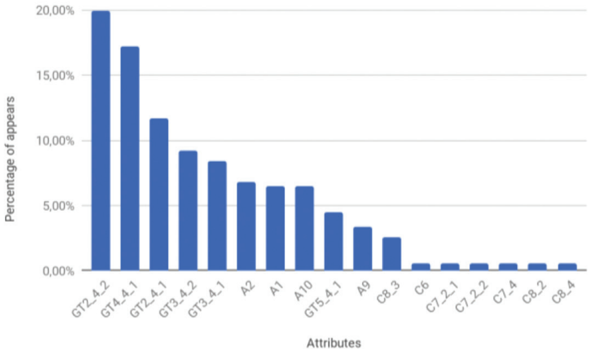

Figure 2: Ranking of variables for leisure expenses

The new Classification model results are: an accuracy rate of 74.20%, a Type I error of 2.27%, a Type II error of 23.53% and a root mean square error of 0.51 with only 17 variables. Table 4 shows the full confusion matrix.

3.2 Spending on transport

Classification models (that consider all attributes) obtain an accuracy rate of 61.65%, a Type I error of 20.71%, a Type II error of 17.63% and root mean square error of 0.93. Table 5 shows the full confusion matrix.

Table 4: Confusion Matrix for new leisure classification model.

	Very Low: 1–60	**High: >80**
Very Low: 1–60	727	59
High: >80	611	1200

Table 5: Confusion Matrix for original transport classification model.

	Very Low: 0–10	**Low: 11–20**	**Lower-Middle: 21–40**	**Higher-Middle: 41–60**	**High: 61–100**	**Very High: > 100**
Very Low:0–10	933	169	52	14	4	2
Low:11–20	210	633	263	29	22	22
Lower-Middle 21–40	3	160	1191	287	129	128
Higher-Middle: 41–60	0	0	177	456	127	26
High:61–100	0	1	76	352	586	139
Very High:> 100	17	5	4	22	176	407

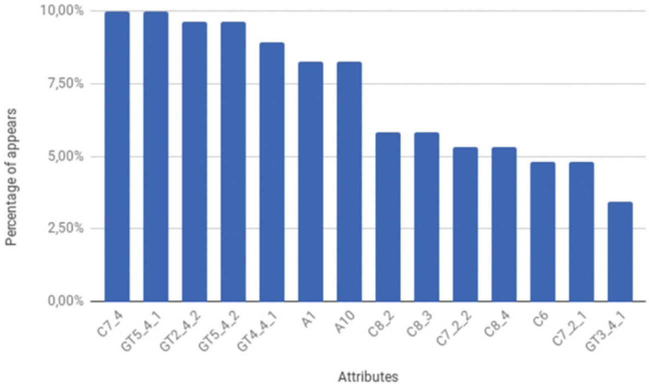

Figure 3: Ranking of variables for transport expenses

The most frequent attributes in the association rules that lead to minimum and maximum spending on transport are: the person who paid the tourist package (C7_4) and the person who paid a rented vehicle (GT5_4_1).

The new Classification model issues are: an accuracy rate of 91.89%, a Type I error of 0.26%, a Type II error of 7.84% and a root mean square error of 0.29 with only 14 variables. Table 6 shows the full confusion matrix.

3.3 Spending on accommodation

The classification model with all attributes obtains an accuracy rate of 83.65%, a Type I error of 8.66%, a Type II error of 7.68% and a root mean square error of 0.61. Table 7 shows the full confusion matrix.

The most frequent attribute in the association rules for spending on accommodation is clearly the reason for the trip (C1). Also, attributes like agency use (C10_1) or advanced booking (C10_2_1).

Table 6: Confusion Matrix for new transport classification model.

	Very Low: 0–10	**High: >100**
Very Low: 0–10	1015	5
High: > 100	148	719

Table 7: Confusion Matrix for original accommodation classification model.

	Equal to 0	**Low: 1–20**	**Middle: :21–40**	**High: > 40**
Equal to 0	3392	126	115	120
Low: 1–20	2	404	151	13
Middle: 21–40	0	74	665	66
High: > 40	6	11	431	1245

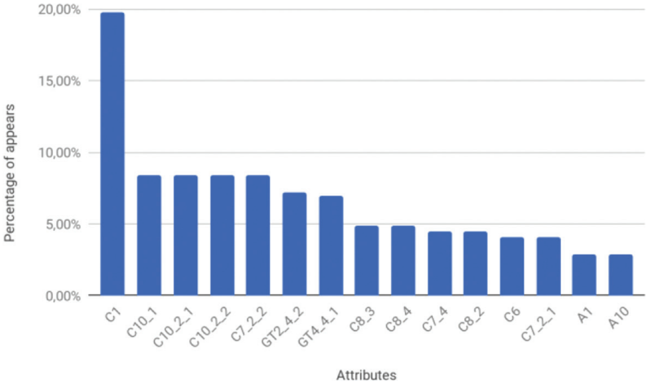

Figure 4: Ranking of attributes for accommodation

The new Classification model results are: an accuracy rate of 97.36%, a Type I error of 2.48%, a Type I error of 0.17% and a root mean square error of 0.16 with only 15 variables. Table 8 shows the full confusion matrix.

3.4 Total spending

Classification models that consider all attributes obtain an accuracy rate of 55.20%, a Type I error of 21.14%, a Type II error of 23.66% and root mean square error of 0.84 (Table 9).

The most frequent attributes are the following (Fig. 5):

Some examples for the most frequent attributes in the association rules that lead to minimum and maximum total expenses are: who paid the transport (GT2_4_2) or how many nights their stay is (A10).

Table 8: Confusion Matrix for new accommodation classification.

	Equal to 0	**High: > 40**
Equal to 0	3392	120
High: > 40	8	1324

Table 9: Confusion Matrix for spending

	Very Low: 1–60	**Low: 61–100**	**Middle: 101–130**	**High: 131–200**	**Very High: > 200**
Very Low: 1–60	666	180	8	3	1
Low: 61–100	554	1292	353	192	62
Middle: 101–130	14	149	175	122	25
High: 131–200	1	171	542	1009	496
Very High: > 200	0	1	11	171	623

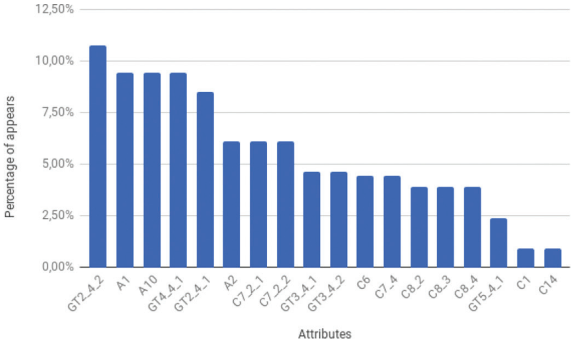

Figure 5: Rank by apriori attribute selector for Total spending

The new Classification model issues are: an accuracy rate of 85.38%, a Type I error of 5.81%, a Type II error of 8.80% and a root mean square error of 0.38 with only 18 variables. Table 10 shows the full confusion matrix.

Below, Table 11 summarizes the above results so as to compare the classification task improvement when completed.

Table 10: Confusion Matrix for new total spending classification model.

	Very Low:1–60	**High: > 200**
Very Low: 1–60	1020	142
High: > 200	215	1065

Table 11: Summary of different models

	ORIGINAL CLASSIF. MODEL					NEW CLASSIF. MODEL				
	(considering full set attributes)					(considering selected attributes)				
Spending on…	**var**	**Accuracy**	**SME**	**Type I error (FP)**	**Type II error (FN)**	**var**	**Accuracy**	**SME**	**Type I error (FP)**	**Type II error (FN)**
Leisure	52	42.95%	1.23	23.35%	33.70%	17	74.20%	0.51	2.27%	23.53%
Transport	52	61.65%	0.93	20.71%	17.63%	14	91.89%	0.29	0.26%	7.84%
Accommodation	52	83.65%	0.61	8.66%	7.68%	15	97.36%	0.16	2.48%	0.17%
Total	52	55.20%	0.84	21.14%	23.66%	18	85.38%	0.38	5.81%	8.80%

By using the A Priori Min-Max algorithm, we reduce the number of variables trained by the CART algorithm from 52 to 17 (leisure), 14 (transport), 15 (accommodation) and 18 (total), improving the accuracy associated to each model in all the cases.

4 CONCLUSIONS

In most cases, the results obtained improve considerably when the A Priori Min-Max algorithm (with minimum and maximum values) is used instead of the original attributes CART selection (with Gini index) trained against the whole dataset. Of course the binarization procedure provides such a high confidence ratio.

The main advantage obtained with the A Priori Min-Max method is that it guarantees an ordered methodology that allows the inclusion of a variable according to its relevance, thereby providing more control over the tree creation. Furthermore, the most frequent patterns are extracted in order to observe the evolution of the data.

The final variable sets chosen by the CART and the A Priori Min-Max algorithm have been evaluated by expert public managers in tourism. These agents confirmed that the variables chosen by the A Priori Min-Max algorithm (the most frequent attributes) have a greater strategic interest in most of typical study cases for any type of expense analysed [7], regardless of the particular sample used for the computational experiment.

5 FUTURE RESEARCH LINES

Improvement in accuracy in each case is due to binarization in extreme expenses and not so much to attribute selection. Thus, future studies will proceed to test other methods for attribute selection from datasets with prior binarization of extreme expenses [8].

On the other hand, the CART algorithm itself can be modified in terms of different information gain criteria, which necessarily leads to different attribute selections.

REFERENCES

[1] Rabasa, A., Pérez-Martín, A. & Giner, D., Optimal clustering techniques for the segmentation of tourist spending. Analysis of tourist surveys in the Valencian community (Spain): a case study. *International Journal of Design & Nature and Ecodynamics*, **12**(4), pp. 482–491, 2018.
https://doi.org/10.2495/dne-v12-n4-482-491

[2] 'EGATUR', available at http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177002&menu=ultiDatos&idp=1254735576863 (accessed March 2018)

[3] Breiman, L. (ed), *Classification and regression trees*, Repr. Boca Raton: Chapman & Hall [u.a.], 1998.

[4] Aksoy, S. & Ozbuk, M.Y., Multiple criteria decision making in hotel location: does it relate to postpurchase consumer evaluations?. *Tourism Management Perspectives*, **22**, pp. 73–81, 2017.
https://doi.org/10.1016/j.tmp.2017.02.001

[5] Nilashi, M., Bagherifard, K., Rahmani, M. & Rafe, V., A recommender system for tourism industry using cluster ensemble and prediction machine learning techniques. *Computers & Industrial Engineering*, **109**, pp. 357–368, 2017.
https://doi.org/10.1016/j.cie.2017.05.016

[6] Pronoza, E., Yagunova, E. & Volskaya, S., Aspect-Based Restaurant Information Extraction for the Recommendation System. *Conference: 6th Language and Technology Conference. Lecture Notes in Computer Science*, 9561, pp. 371–385, 2016.

[7] Sinclair, M.T. & Stabler, M., *The Economics of Tourism*. Library of Congress Cataloguing in Publication Data, 2002.

[8] Pol, A.P., Pascual, M.B. & Vazquez, P.C., Robust estimators and bootstrap confidence intervals applied to tourism spending. *Tourism Management*, **27**(1), pp. 42–50, 2006. https://doi.org/10.1016/j.tourman.2004.06.016