International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 1, February 2021, pp. 409~415
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i1.pp409-415

Benchmarking data mining approaches for traveler segmentation

Tamer Uçar, Adem Karahoca
Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, Turkey

Article Info

Article history:

Received Aug 7, 2018
Revised Jun 16, 2020
Accepted Jun 29, 2020

Keywords:

Data mining
Hybrid data mining approach
Machine learning algorithms
Travel planning

ABSTRACT

The purpose of this study is proposing a hybrid data mining solution for traveler segmentation in tourism domain which can be used for planning user-oriented trips, arranging travel campaigns or similar services. Data set used in this work have been provided by a travel agency which contains flight and hotel bookings of travelers. Initially, the data set was prepared for running data mining algorithms. Then, various machine learning algorithms were benchmarked for performing accurate traveler segmentation and prediction tasks. Fuzzy C-means and X-means algorithms were applied for clustering user data. J48 and multilayer perceptron (MLP) algorithms were applied for classifying instances based on segmented user data. According to the findings of this study, J48 has the most effective classification results when applied on the data set which is clustered with X-means algorithm. The proposed hybrid data mining solution can be used by travel agencies to plan trip campaigns for similar travelers.

Corresponding Author:

Tamer Uçar,
Department of Software Engineering,
Faculty of Engineering and Natural Sciences,
Bahcesehir University, Besiktas, Istanbul, 34353, Turkey.
Email: tamer.ucar@eng.bau.edu.tr

1. INTRODUCTION

Data mining is a technique for extracting knowledge from large data sets. It is the combination of statistical and mathematical methods for processing raw data to discover knowledge [1]. Today, data mining methods are used in many topics such as filtering systems, risk analysis management, fraud detection, medicine, e-commerce and many more [2-4]. Tourism domain is one of these areas where different types of data mining solutions can be applied. Tour planners, travel schedule planners and social media-based trip recommenders are examples of possible data mining related works [5].

Tour planners suggest possible visit locations to its users. Location based collaborative filtering approach can be used for this purpose [6]. I. García-Magariño [7] proposed an agent-based tour simulator whereas A. Varfolomeyev, et al. [8] focused on generating recommendations for historical tourism planning, and R. Colomo-Palacios, et al. [9] proposed a context-aware recommender system for mobile devices. Travel schedule planners help its users to build time tables for visit locations by taking time and related constraints into account. F. M. Hsu, et al. [10] combined Engel-Blackwell-Miniard model with Bayesian network. A. Moreno, et al. [11] proposed an ontology-based recommendation system. [12] combined trip planning and scheduling. Content-based filtering and hotel service recommenders are proposed by [13, 14].

Social media-based trip recommenders propose items based on information retrieved from sources like geo-tagged photos of travelers [15]. Y. Sun, et al. [16] Used geo-tagged image data for road-based recommendations. Extracted trip behaviors of users from geo-tagged photos [17]. Identified tourist hot spots

using photos from social networks [18]. A. Majid and J. Han [19, 20] Proposed similar approaches for obtaining and personalizing travel locations. To carry out these tasks a reliable data mining framework is required [21-23].

An expert system which is designed for the tasks above mostly relies on a recommender engine. Basically, a recommender engine tries to propose similar items to a target user or user group [24]. To achieve this goal, system tries to generate a rating value. Possible items are matched with users based on the generated rating score. Rating score computation can be carried out in different ways. Most systems use target user's profile and previous user behavior data for this task.

In this study, selected data mining methods were tested and benchmarked on a traveler data set to propose a possible hybrid data mining approach for getting accurate travel recommendations. Neural network-based and tree-based data mining methods were combined with Fuzzy C-means and X-means clustering algorithms to assess the data mining model pair which generates the highest prediction correctness. The rest of the paper is organized as follows: Section 2 describes the data gathering process and includes a background of the data mining algorithms used in this study. Section 3 presents the obtained results and Section 4 contains conclusion of this study.

2. RESEARCH METHOD

Different types of machine learning algorithms were tested on a real-world traveler data set which contains information about flight and hotel booking transactions of travelers. Detailed definition about this data set and applied algorithms are defined in the following subsections.

2.1. Data gathering and processing

The raw data set was retrieved from a travel agency. It included transactions of 26,886 flight bookings and 4,367 hotel bookings. After finding flight and hotel records that customers booked for the same trip, 317 matching records were collected. Removing the identity columns from this data set yielded 14 attributes. Table 1 lists these initial data set attributes and their descriptions.

Further data set analysis revealed that "departure location" and "returning location (to)" attributes were containing the same set of values. Because of this fact, "departure location" was removed from data set. Values of "returning location (from/to)" attributes were discretized according to regions. Table 2 lists possible regions and their numeric codes. Values for the flight and hotel cost attributes were discretized into six groups according to customers' expenses. Table 3 and Table 4 show discretized cost groups.

"Departure date" and "returning date" attributes were used for computing each transaction's travel season and travel duration values. "Days in hotel" attribute was removed because its values were same with travel duration. And ticket class attributes were removed because 97% of ticket class values were from the same ticket type.

After deriving two new attributes (travel season, travel duration), removing redundant fields and discretizing data set, 10 attributes were collected and preprocessed for data mining algorithms. Table 5 lists the final data set attributes and their descriptions. The final data set was used for training and testing data mining models. 66% of data was used for training and 34% was used for testing models.

Table 1. Initial data set attributes

Attribute	Description
Gender	Passenger's gender.
Departure date	Starting date of travel.
Departure location	Location which the passenger is leaving form.
Arrival location	Location which the passenger is arriving to.
Departure airline	Airline company for departure flight.
Departure flight class	Ticket class for departure flight.
Returning date	Ending date of travel.
Returning location (from)	Location which the passenger is returning from.
Returning location (to)	Location which the passenger is returning to.
Returning airline	Airline company for returning flight.
Returning flight class	Ticket class for returning flight.
Flight cost	Flight's cost.
Days in hotel	Number of days stayed in hotel.
Hotel cost	Hotel's cost.

Table 2. Region codes

Code	Description
1	Northern Europe
2	Southern Europe
3	Eastern Europe
4	Western Europe
5	Central Europe
6	Balkans
7	Middle East
8	Northern Asia
9	Southern Asia
10	Eastern Asia
11	Western Asia
12	Central Asia
13	Africa
14	America
15	Australia
16	(Turkey) Marmara Region
17	(Turkey) Black Sea Region
18	(Turkey) Central Anatolia Region
19	(Turkey) Southeastern Anatolia Region
20	(Turkey) Aegean Region
21	(Turkey) Eastern Anatolia Region
22	(Turkey) Mediterranean Region

Table 3. Flight cost groups

Code	Description
1	< 200
2	201 – 400
3	401 – 700
4	701 – 1400
5	1401 – 3000
6	4000 +

Table 4. Hotel cost groups

Code	Description
1	< 350
2	351 – 700
3	701 – 1000
4	1001 – 1500
5	1501 – 2500
6	2500 +

Table 5. Final data set attributes

Attribute	Description
Gender	Passenger's gender.
Travel duration	Duration of travel in days.
Season	Season of travel.
Arrival location	Location which the passenger is arriving to.
Departure airline	Airline company for departure flight.
Returning location (from)	Location which the passenger is returning from.
Returning location (to)	Location which the passenger is returning to.
Returning airline	Airline company for returning flight.
Flight cost	Flight's cost.
Hotel cost	Hotel's cost.

2.2. Clustering and classification algorithms

Various clustering and classification algorithms were executed to build prediction models using the described traveler data set. Brief descriptions of these approaches are listed below:

a) *Multilayer perceptron (MLP):* MLP is a classification algorithm based on feed-forward artifical neural network models. It employs backpropogation for training the network [1].
b) *J48:* J48 is the Java implementation of C4.5 decision tree algorithm which is based on ID3. Information entropy is used by this approach while constructing the decision tree model [1, 25].
c) *Fuzzy C-means clustering (FCM):* FCM is a soft clustering algorithm. Unlike hard clustering methods, each point in a data set has a degree of belonging to clusters [26, 27].
d) *X-means clustering (XM)*: XM can be summarized as an improved version of the K-means clustering algorithm it provides self-estimation of the number of clusters for a given data set [28].

2.3. Comparing algorithms

Non-binary confusion matrix is used as the primary tool for computing classification metrics of data mining models. Row indices of the matrix show actual values and column indices show predicted values for a classification task. Based on the values of a confusion matrix, various metrics can be computed for comparing data mining algorithms. Most common metrics are true positive (TP), false negative (FN),

false positive (FP), true negative (TN), true positive rate (TPR), true negative rate (TNR), precision, correctness and root mean squared error (RMSE).

TP is the number of positive examples correctly predicted by the classification model. FN is the number of positive examples wrongly predicted as negative whereas FP is the number of negative examples wrongly predicted as positive and TN is the number of negative examples correctly predicted by the classification model. TPR (recall) is the fraction of positive examples predicted correctly and TNR (specificity) is the fraction of negative examples predicted correctly by the classification model. Precision is the ratio of TP instances by the total number of TP and FP instances. Correctness is the percentage of correctly classified instances. Root mean squared error (RMSE) is a metric which is computed for assessing differences between actual and predicted instances.

2.4. Model training

WEKA [29] and MATLAB [30] tools were used for running the clustering and classification algorithms. Comparison metrics which are described above are computed for each prediction model and obtained results are discussed in the next section.

3. RESULTS AND ANALYSIS

Final version of the traveler data set was segmented into four to eight clusters using X-means and Fuzzy C-means clustering algorithms. This process yielded ten differently segmented versions for the same data set. Using each differently segmented set, J48 and MLP prediction models were generated. The most accurate model among these models can be used for classifying the corresponding segmentation of a new traveler instance in an accurate way. Table 6 lists recall, specificity, precision, correctness and RMSE values of each prediction model.

Table 6. Benchmarking prediction models

Classifier	Clusterer	Cluster Size	Recall	Specificity	Precision	Correctness	RMSE
J48	XM	5	0.98	0.99	0.98	98.15	0.09
J48	FCM	4	0.96	0.98	0.96	96.30	0.14
MLP	FCM	4	0.96	0.99	0.96	96.30	0.13
MLP	XM	8	0.96	0.99	0.97	96.30	0.08
J48	XM	4	0.95	0.98	0.96	95.37	0.15
J48	XM	8	0.95	0.99	0.95	95.37	0.11
MLP	XM	5	0.95	0.99	0.96	95.37	0.14
MLP	XM	6	0.95	0.99	0.96	95.37	0.11
MLP	XM	7	0.95	0.99	0.96	95.37	0.09
J48	XM	6	0.94	0.98	0.95	94.44	0.14
MLP	FCM	5	0.94	0.99	0.95	94.44	0.13
MLP	XM	4	0.94	0.98	0.95	94.44	0.16
J48	XM	7	0.94	0.98	0.94	93.52	0.11
J48	FCM	5	0.93	0.98	0.93	92.59	0.16
MLP	FCM	6	0.93	0.98	0.93	92.59	0.14
MLP	FCM	8	0.93	0.99	0.93	92.59	0.13
J48	FCM	8	0.91	0.98	0.90	90.74	0.14
J48	FCM	7	0.90	0.98	0.89	89.81	0.16
MLP	FCM	7	0.90	0.98	0.90	89.81	0.14
J48	FCM	6	0.89	0.98	0.90	88.89	0.18

According to the obtained experimental results shown in Table 6, J48 has the best correctness, precision and recall scores when it is applied on the data set clustered into five clusters using X-means algorithm. MLP generates the highest specificity and lowest RMSE values when it is applied on the data set clustered into eight clusters using X-means algorithm. The best score for each metric was obtained by the X-means clustering algorithm. Table 7 shows the decision tree paths for the J48 and X-means method combination which has the top correctness score.

According to the listed results in Table 7, J48 model generated 11 different tree paths. Each path can be mapped as a decision rule for a specific type of a customer. Based on the listed paths, characteristics of each cluster can be defined as follows:

1) Cluster 1 represents male or female passengers whose preferred returning location is within location codes from 1 to 14 and preferred returning airline is within company codes from 10 to 77.
2) Cluster 2 represents four different types of passengers:

a. Male or female passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is above 1000 TL.
b. Male passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is between 701 TL and 1000 TL. These passengers prefer travelling in summer or fall seasons.
c. Male passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is no more than 700 TL.
d. Female passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is between 351 TL and 1000 TL.

3) Cluster 3 represents three different types of passengers:
   a. Male passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is between 701 TL and 1000 TL. These passengers prefer travelling in spring or winter seasons.
   b. Female passengers whose preferred returning location is within location codes from 15 to 22 and hotel cost is no more than 350 TL.
   c. Male or female passengers whose preferred returning location is within location codes from 1 to 14 and preferred returning airline is within company codes from 1 to 9 and preferred departure airline is any company other than the company with code 1 and hotel cost is no more than 700 TL. These passengers prefer travelling in summer or fall seasons.
4) Cluster 4 represents male or female passengers whose preferred returning location is within location codes from 1 to 14 and preferred airline is within company codes from 1 to 9 and hotel cost is above 700 TL. These passengers prefer travelling in summer or fall seasons.
5) Cluster 5 represents two different types of passengers:
   a. Male or female passengers whose preferred returning location is within location codes from 1 to 14 and preferred returning airline is within company codes from 1 to 9 and preferred departure airline is the company with code 1 and hotel cost is no more than 700 TL. These passengers prefer travelling in summer or fall seasons.
   b. Male or female passengers whose preferred returning location is within location codes from 1 to 14 and preferred returning airline is within company codes from 1 to 9. These passengers prefer travelling in winter or spring seasons.

Table 7. J48 decision tree paths

Path #	Path Rule
Path 1	If “Returning location (from)” > 14 and “Hotel Cost” > 3 Then output is Cluster 2
Path 2	If “Returning location (from)” > 14 and “Hotel Cost” <= 3 and “Gender” > 0 and “Hotel Cost” > 2 and “Season” > 2 Then output is Cluster 2
Path 3	If “Returning location (from)” > 14 and “Hotel Cost” <= 3 and “Gender” > 0 and “Hotel Cost” > 2 and “Season” <= 2 Then output is Cluster 3
Path 4	If “Returning location (from)” > 14 and “Hotel Cost” <= 3 and “Gender” > 0 and “Hotel Cost” <= 2 Then output is Cluster 2
Path 5	If “Returning location (from)” > 14 and “Hotel Cost” <= 3 and “Gender” <= 0 and “Hotel Cost” > 1 Then output is Cluster 2
Path 6	If “Returning location (from)” > 14 and “Hotel Cost” <= 3 and “Gender” <= 0 and “Hotel Cost” <= 1 Then output is Cluster 3
Path 7	If “Returning location (from)” <=14 and “Returning Airline” > 9 Then output is Cluster 1
Path 8	If “Returning location (from)” <=14 and “Returning Airline” <= 9 and “Season” > 2 and “Hotel Cost” > 2 Then output is Cluster 4
Path 9	If “Returning location (from)” <=14 and “Returning Airline” <= 9 and “Season” > 2 and “Hotel Cost” <= 2 and “Departure Airline” > 1 Then output is Cluster 3
Path 10	If “Returning location (from)” <=14 and “Returning Airline” <= 9 and “Season” > 2 and “Hotel Cost” <= 2 and “Departure Airline” <= 1 Then output is Cluster 5
Path 11	If “Returning location (from)” <=14 and “Returning Airline” <= 9 and “Season” <= 2 Then output is Cluster 5

4. CONCLUSION

This study presents and compares detailed model performances of different data mining algorithms executed on a real-world traveler data set. Based on the obtained results, J48 and X-means algorithm combination has the best prediction performance in terms of given classification metrics. This hybrid data mining method combination can be used for predicting possible trip destinations based on behaviors of similar users. The prediction result can support decision-making process of travel agencies while preparing campaigns. Alternatively, it can be a part of a travel system where possible trip opportunities can be proposed to similar users. Including more classification and clustering algorithms to this apprach can be modeled as a part of a future study.

REFERENCES

[1] I. H. Witten, et al., "Data Mining: Practical machine learning tools and techniques," *The Morgan Kaufmann Series in Data Management Systems*, 2011.

[2] K. A. Almohsen and H. Al-Jobori, "Recommender systems in light of big data," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 6, pp. 1553-1563, 2015.

[3] M. S. Bhatt and T. P. Patalia, "Indian monuments classification using support vector machine," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 4, pp. 1952-1963, 2017.

[4] M. Nasiri, et al., "Fetal electrocardiogram signal extraction by ANFIS trained with PSO method," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 2, no. 2, pp. 247-260, 2012.

[5] J. Borràs, et al., "Intelligent tourism recommender systems: A survey," *Expert Systems with Applications*, vol. 41, no. 16, pp. 7370-7389, 2014.

[6] W. S. Yang and S. Y. Hwang, "iTravel: A recommender system in mobile peer-to-peer environment," *Journal of Systems and Software*, vol. 86, no. 1, pp. 12-20, 2013.

[7] I. García-Magariño, "ABSTUR: An agent-based simulator for tourist urban routes," *Expert systems with applications*, vol. 42, no. 12, pp. 5287-5302, 2015.

[8] A. Varfolomeyev, et al., "Smart space based recommendation service for historical tourism," *Procedia Computer Science*, vol. 77, pp. 85-91, 2015.

[9] R. Colomo-Palacios, et al., "Towards a social and context-aware mobile recommendation system for tourism," *Pervasive and Mobile Computing*, vol. 38, no. 2, pp. 505-515, 2017.

[10] F. M. Hsu, et al., "Design and implementation of an intelligent recommendation system for tourist attractions: The integration of EBM model, Bayesian network and Google Maps," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3257-3264, 2012.

[11] A. Moreno, et al., "Sigtur/e-destination: ontology-based personalized recommendation of tourism and leisure activities," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 1, pp. 633-651, 2013.

[12] P. Aksenov, et al., "Toward personalised and dynamic cultural routing: a three-level approach," *Procedia Environmental Sciences*, vol. 22, pp. 257-269, 2014.

[13] H. S. Chiang and T. C. Huang, "User-adapted travel planning system for personalized schedule recommendation," *Information Fusion*, vol. 21, no. 1, pp. 3-17, 2015.

[14] N. Silamai, et al., "TripRec: trip plan recommendation system that enhances hotel services," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pp. 412-420, 2017.

[15] Z. Xu, et al., "Topic based context-aware travel recommendation method exploiting geotagged photos," *Neurocomputing*, vol. 155, pp. 99-107, 2015.

[16] Y. Sun, et al., "Road-based travel recommendation using geo-tagged images," *Computers, Environment and Urban Systems*, vol. 53, pp. 110-122, 2015.

[17] I. R. Brilhante, et al., "On planning sightseeing tours with TripBuilder," *Information Processing & Management*, vol. 51, no. 2, pp. 1-15, 2015.

[18] J. C. García-Palomares, et al., "Identification of tourist hot spots based on social networks: A comparative analysis of European metropolises using photo-sharing services and GIS," *Applied Geography*, vol. 63, pp. 408-417, 2015.

[19] A. Majid, et al., "A system for mining interesting tourist locations and travel sequences from public geo-tagged photos," *Data & Knowledge Engineering*, vol. 95, pp. 66-86, 2015.

[20] J. Han and H. Lee, "Adaptive landmark recommendations for travel planning: personalizing and clustering landmarks using geo-tagged social media," *Pervasive and Mobile Computing*, vol. 18, pp. 4-17, 2015.

[21] J. Han, et al., "Data mining: concepts and techniques," *The Morgan Kaufmann Series in Data Management Systems*, 2011.

[22] T. Uçar, et al., "NTRS: A New Travel Recommendation System Framework by Hybrid Data Mining," *International Journal of Mechanical Engineering and Technology*, vol. 10, no. 01, pp. 935-946, 2019.

[23] A. Karahoca and D. Karahoca, "GSM churn management by using fuzzy c-means clustering and adaptive neuro fuzzy inference system," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1814-1822, 2011.

[24] L. Lü, et al., "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1-49, 2012.

[25] J. R. Quinlan, "C4. 5: programs for machine learning," *Morgan Kaufmann Series in Machine Learning*, 1992.

[26] J. C. Bezdek, et al., "Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines," *SIAM Journal on Applied Mathematics*, vol. 40, no. 2, pp. 339-357, 1981.

[27] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32-57, 1973.

[28] D. Pelleg and A. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," *Proceedings of the 17th International Conference on Machine Learning*, vol. 1, pp. 727-734, 2000.

[29] I. H. Witten, et al., "The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques," *Morgan Kaufmann*, 2016.

[30] S. N. Sivanandam, et al., "Introduction to fuzzy logic using MATLAB," Berlin, *Springer*, 2007

BIOGRAPHIES OF AUTHORS

Dr. Tamer Uçar holds a PhD in Computer Engineering. He is interested in software development, data mining, recommender systems, medical data analysis and big data. He has published articles about data mining applications in various fields. He is currently working for Bahçeşehir University Software Engineering Department as a full-time faculty member.

Dr. Adem Karahoca holds a PhD in Software Engineering. He is interested in human–computer interaction, web-based education systems, data mining, big data, and management information systems. He has published articles at prestigious journals about use and data mining applications of business information systems in health, tourism, and education.