E-ISSN: 2676-6167 | P-ISSN: 2538-5100

Journal of Applied Research on Industrial Engineering

www.journal-aprie.com

J. Appl. Res. Ind. Eng. Vol. 10, No. 1 (2023) 11–24.

Paper Type: Research Paper

Analyzing Behavioral Patterns of Bus Passengers Using Data Mining Methods (Case Study: Rapid Transportation Systems)

Amir Daneshvar[1], Fariba Salahi[2,*], Maryam Ebrahimi[1], Bijan Nahavandi[3]

[1] Department of Information Technology Management, Faculty of Management, Electronic Branch, Islamic Azad University, Tehran, Iran; daneshvar.amir@gmail.com; mar.ebrahimi@gmail.com.

[2] Department of Industrial Management, Faculty of Management, Electronic Branch, Islamic Azad University, Tehran, Iran; salahi_en@yahoo.com.

[3] Department of Industrial Management, Faculty of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran; bnahavandi@gmail.com.

Citation:

Daneshvar, A., Salahi, F., Ebrahimi, M., & Nahavandi, B. (2023). Analyzing behavioral patterns of bus passengers using data mining methods (case study: rapid transportation systems). *Journal of applied research on industrial engineering, 10*(1), 11-24.

Received: 26/08/2021 Reviewed: 01/10/2021 Revised: 09/11/2021 Accepted: 10/01/2022

Abstract

The aim of analyzing passengers' behavioral patterns is providing support for transportation management. In other words, to improve services like scheduling, evacuation policies, and marketing, it is essential to understand spatial and temporal patterns of passengers' trips. Smart Card Automated Fare Collection System (SCAFCS) makes it possible to utilize data mining tools for the purpose of passengers' behavioral pattern analysis. The specific goal of this research is to obtain functional information for passenger's behavioral pattern analysis in city express bus which is called BRT, and classification of passengers to improve performance of bus fast transportation system. Additionally, it is attempted to predict usage and traffic status in a line through predicting passenger's behavior in a bus line. This paper applies smart card data to provide combinational algorithms for clustering and analysis based on data mining. To this end, we have used a combination of data mining methods and Particle Swarm Optimisation (PSO) algorithm and leveraged multivariate time series prediction to estimate behavioral patterns. Results show that price and compression ratio features are the most influencing features in the separability of transportation smart card data. According to the Pareto front, four features: card identification number, bus identification number, bus line number, and charge times influence clustering criteria.

Keywords: Smart card, E-ticket, Artificial intelligence, Particle swarm optimisation, Behavioral pattern.

1 | Introduction

Urban society's health depends on macro policy makings in urban planning and especially transportation planning; thus, paying attention to public transportation is essential in daily life and is considered necessary for improving the quality of life in urban areas. Paying attention to this issue would reduce most people's problems significantly [1].

Transportation is one of the main milestones of urban development, used to transport people and goods between different spaces and locations. By extension and expansion of metropolises and increasing demand for fast and cost-efficient intercity transportation, the need for efficient urban systems that can transport a large volume of passengers resulted in the emergence of metro and urban express buses, which are called BRT. Currently, this system is considered one of the most

Corresponding Author: salahi_en@yahoo.com

https://doi.org/10.22105/jarie.2021.302560.1373

critical servers. According to this point, we should take care that if ignoring passengers' short-term and long-term behavioral patterns and not serving them efficiently, we would face a reduction in using these services. Subsequently, there is the possibility of using personal cars, urban traffic, and environmental problems [2]. To establish a comprehensive and appropriate transportation system, designers should predict transportation demand by altering system characteristics and how people use it. Trip demand models, significantly predicting trip characteristics and application of transportation facilities under different scenarios, are proposed [3]. To realistic predictions in trip demand modeling, we need to use behavioral features. Nowadays, these needs have become more serious, since long-term guidelines have been replaced by short-term policies like work scheduled and remote communication and trip spatial-temporal patterns of passengers.

The proper understanding of separate passengers' behaviors helps transportation officials evaluate their current services and adjust their strategies to improve efficiency. For example, suppose that a company provides tickets applicable for one month; this would let passengers ensure consistency of their fares for one month. Evaluating their behavior would help estimate traffic and income charges. In most transportation devices, issues like not paying the cost or stealing exist. We should locate these inappropriate behaviors by passenger's behavioral analysis. One of the main problems in this area is the large number of outliers that increase errors in the clustering section. Also, in practice, it would result that the algorithm would not converge by applying a large amount of data. Another problem is the existence of multiple start points, which result in a difference in obtained results from the same data.

To overcome these problems, we should simultaneously improve clustering performance and the correct data set management. Thus, we should apply appropriate preprocessing, including correlation reduction and elimination of outliers, and simultaneously, we should use proper models and powerful clustering methods. Hence in this paper, to achieve this goal utilizing smart cards data, we have proposed efficient combined algorithms for clustering and analysis based on data mining. Generally, we could help to improve transportation system performance and efficient management in this method, utilizing operations like data cleaning, data labeling and clustering, classification, multiobjective optimisation, and intuitive analysis.

The current study has utilized data mining and artificial intelligence to determine how we could predict passengers' behavioral patterns using urban express buses operating smart card (e-ticket) data?

2 | Literature Review

A review of internal and external research shows that planning on efficient transportation systems is essential for countries. As proposed in [4], a linear programming model minimize the set of line set up costs, maintenance, repair costs, fuel consumption, and the time of bus entrance to the station. The proposed model is simulated in GAMS software, and sensitivity analysis has been performed on that. According to the proposed model's high complexity and computational time, a combinational meta-heuristic method based on Particle Swarm Optimisation (PSO) and Ant Colony Optimization (ACO) was developed to solve it and simulated in MATLAB software. Results show that fuel cost and spatial distance between stations have the most influence on the welfare of passengers. Also, in terms of model complexity and solution time, the number of stations has more effect than the number of lines and buses.

The study of [5] is about studying the nature and dimensions of express buses' full passenger line using analyzing strengths and weaknesses, opportunities and threats, and evaluating the feasibility of concurrent usage of bus lines. The performance of Vahed Company bus drivers has been evaluated in [6]. The results have shown a meaningful correlation between danger and actual accidents of drivers with error, violations, age, and license type of drivers.

Research on the topic of dynamic modeling of urban transportation systems has been performed in [7]. Its purpose was to provide a dynamic model for urban transportation systems regarding a stable urban transportation system. Using this model, they have understood the complex and dynamic nature of the urban transportation systems and evaluated the influence of policies. In this paper, they have utilized the systems dynamics approach to propose a comprehensive and integrated model for the Tehran urban transportation system regarding the concept of the stable urban transportation system and evaluated the system's dynamic behavior. Furthermore, the simulated model provides policies for an increase in urban transportation systems' efficiency and improve parameters of urban traffic. This model has been analyzed using Vensim software and Tehran city data. Results showed that combined components of proposed policies greatly influence enhancing the efficiency of the transportation system and statutes of Tehran city traffic parameters.

In recent years, many types of research have been performed on using smart card data in public transportation. Obtained data of smart cards have been used in many applications like analyzing demand amount and planning and others [8]. It has been used in [9], [10], a smart card Automatic Fare Collection (AFC) system to define various consumers and measure their trip habits; analyses were based on patterns in days, weeks, and seasons. Also, in many works, they have used the clustering method as a data mining method in e-ticket information classification. Among these, the scaled k-means++ enhanced method has been used for data with large volumes and a high changing rate [11]. The stages of data mining in this paper are mention below:

I. Data filtering.
II. Data clustering (using k-means and HAC algorithms).
III. Describing groups.

In this work, we have used features of time, date, and type of card and correspond line direction (25452) of a smart card; after clustering, we obtained four groups of clusters. Results have shown that peak usage of these individuals is in business center closure, and these individuals usually use transportation systems to round trips between work and home [10]. In [9], daily data of 4.47 million cards are used for various analyses. In these works, only three-parameter of the card number, fare type, and date and time. They have used k-means clustering to classify data. According to the data vision and performed distance calculations, the number of clusters has been set to three. Results show that the first cluster has a transaction peak on a typical day, and we could say that registered trips in this cluster have extended during fay. The second cluster corresponds to morning trips; since post-morning transactions were very low, many transactions have been reported between 9 to noon. In the third cluster, most trips have registered in post morning and evening. Also, the transaction percentage of the three clusters shows that on working days (Monday to Friday), individuals in the first cluster had the most transactions. On weekend days, individuals in the second and third clusters had the most transportation system usage.

In [11], smart card data and GPS are used to provide a method for estimating the original destination of the OD matrix of Santiago city public transportation. Considering get-off time and location information, they have used two datasets of one week and in different periods to estimate more than 80% of transactions. In [12], a method is proposed based on Markoff save and Bayesian decision-making tree to infer stop stations of bus passengers. This research has to utilize smart card data and data received from GPS installed in buses. In [13], a method is proposed for data extraction to monitor the real-time performance of transportation systems based on smart card data and GPS data.

In [14], a method is proposed for determining significant places based on users' mobile phones (active android). Detection of inappropriate behaviors somehow is an outlier's detection problem. Many methods have been proposed to determining these outliers, which we could mention various techniques based on classification, distance-based on clustering, and others among them. In the following, we have classified performed methods in three main classes, and we will generally evaluate works in these three groups.

2.1 | Strategy Level Studies

In this section, performed works are proportional to long-term planning. Many researchers suggest that we should utilize smart card information's in long-term planning. For example, the study of [10] has mentioned that analysis of smart card data results in a better understanding of users' behaviors. In *Table 1*, a review of performed studies at the strategic level has been presented. The table shows that most of these works are on users' properties and classification without having users' previous private information. In all performed works using smart card data, knowledge of the date, time, and card number are available. We could calculate adequate statistics using them. Of course, since usually information of user person is not available, we will use traditional methods like surveys to compensate for these defects. As it is mentioned earlier, we could use smart cards to measure users' loyalty.

Table 1. Studies that used smart card data for strategic transportation planning.

Reference	Data	Type of Analysis or Data Application	Advantages of Obtained Results
Agard et al. [9]	Data of stations, time, location, and type of card	Definition of regular users and measuring their trip habits. analysis using daily, weekly, or seasonally variations	A better understanding of user behaviors
Alfred Chu et al. [15]	Estimated get-off points and type of card	Temporal-spatial image of the network. Driver Assistance Benefits Implication (DABI)	Improve network lines' geometry and stations schedule. Adopting transportation networks to users demands
Park et al. [16]	Historical data	Estimating of future behavioral trends. Creating future demand matrix	Long-term service adoption Network expansion and predicting requirements of network expansion
Trépanier and Morency [17], Trépanier et al. [18], [19], Trépanier and Vassiviere [20]	Trips transactions using the card. Dates of begin and end of trips	Computing intervals in which users have used cards	Modeling user's loyalty using smart card

This paper will propose an efficient combinational algorithm in clustering and analysis based on smart card data. A novel aspect of this work is utilizing a combination of data mining and artificial intelligence methods to obtain a maximum of useful information and evaluate results obtained from using multivariate time series prediction to estimate behavioral patterns, which has not been done previously. Most of the presented works have used single and straightforward algorithms to perform data mining (evaluation of outliers, clustering, and optimizing). Also, we attempt to predict the status of usage and traffic on a bus line, using passenger behavior prediction in that line.

2.2 | Comparison of Studies

A classification model in order to detect unusual events in the daily life behavior patterns of users of mobile phones using pattern recognition is developed in [21]. The machine learning model was used in [22] to analyze online learning behavior of students. In [23], artificial intelligence approaches useful for behavioral modeling are reviewed among which machine-learning methods sound to be more effective. In [24], behavior analysis with the use of machine learning was introduced.

Table 2. Comparison of studies.

Reference	Pattern Recognition	Machine Learning	Simulation
Ahn and Han [21]	✓		
Yan and Au [22]		✓	
Osoba and Davis [23]		✓	
Ceja [24]		✓	
The present study	✓	✓	✓

Daneshvar et al.| J. Appl. Res. Ind. Eng. 10(1) (2023) 11-24

Recently, some studies have been published ablout forecasting through using Artificial Neural Networks (ANNs). For example, Aliahmadi et al. [25] compared the use of traditional methods and neural networks in time series forecasting. The study of [26] used trend analysis on the time series data and regression analysis to find the cause-and-effect relation among variables. In [27], ANN approaches are applied in change point analysis.

3 | Research Methodology

First, in this section, we have provided a flowchart of the proposed method. As mentioned, to manage and analyze the behavior of passengers in a transportation system, we could utilize smart card data to have a helpful efficiency. A way to achieve this goal is through data mining techniques. Generally, in these methods, we help improve transportation systems' performance and efficient management, using operations like data cleaning, labeling and clustering of data, classification, multi-object optimisation, and intuitive analyses. Finally, we could consider the data mining model the same as the model shown in *Fig. 1*. In this paper, we use several stages, which have been mentioned below:

3.1 | Data Preprocessing

In this stage, we eliminate the noises of data using a median filter. We will convert properties like time, date, and card numbers to integer numbers in data preprocessing to facilitate clustering. Additionally, we have normalized data to reach meaningful results in neural networks.

3.2 | Dimension Reduction and Feature Selection

In the first layer, we have performed a feature selection procedure. We have to utilize multi-object optimisation methods to perform unsupervised feature selection. The proposed method for this layer is to use a multi-object PSO algorithm as a multi-object optimisation method. Proposed object functions are the number of features η_f and statistical validation criteria of space. We could show that the value of these criteria would be reduced with a reduction in various data. The main stages of the Multi Objective Particle Swarm Optimisation (MOPSO) algorithm have mentioned below:

1. Generation of primary population and initialization of speed and location of each particle (in initialization, we will set particle speed vector equal to zero, and we will select location vector randomly).
2. Calculation of cost functions for particle.
3. Finding non-dominant members of the population and storing them in the archive.
4. Generation of hypercubes in object space and positioning them so that these hypercubes represent coordination system and the cost function of particle represent coordination of each particle.
5. Each of the particle randomly chooses a leader from the archive and move toward it.

If the termination condition is not meet, go to Stage 3 and otherwise terminate the algorithm.

Regarding feature selection, we have encoded each feature's existing with "1" and the lack of that feature with "0". Hence, in the application of the PSO algorithm, feature selection of particle' components location vector is in terms of 0 and 1. In Coello et al. [28], a jump based on uniform distribution has been proposed to improve the algorithm's performance. In this study, we suggest including jump and intersection operators in the BMOPSO algorithm. The procedure is that we apply the jump operator on 20% of particle with the most cost, and (after excluding the previous 20%) we apply the intersection operator on 60% of particle with the most price. In the intersection operator, we randomly choose two-point of two-particle vector (parents), and then we exchange values of two particle between these intersection points.

In the proposed method, a one-bit jump means converting 0 to 1 and, inversely, a jump probability of pm. We randomly choose a number between 0 and 1, and then if this number is lower than pm, we change the corresponding bit, and otherwise, we will not change the bit. We apply jump on bits forming particle independently. Using intersection and jump operators has resulted in good improvement in the proposed algorithm of multi-object optimisation.

In the next stage, we pass selected features along with the number of clusters to the clustering algorithm, and then we calculate space statistics using these separated data. In this section, we have used the k-means algorithm as a clustering algorithm. In multi-object PSO algorithm or BMOPSO, each particle contains $\eta_f + k_c$ bit in which η_f is the number of features and k_c is the number of clusters, and have encoded in 4-bit form $k \in \{2, \dots, 17\}$.

3.3 | Clustering of Smart Card Data

Matrix X is original data with a dimension of $N \times d$, in which N is the number of observations and d is the number of features. The output of the first layer is matrix X with dimensions of $N \times R$, in which R≤d is selected features. The second layer is responsible for clustering data in the K cluster. The proposed method in this layer is a combination of Rough K-means clustering algorithm and PSO optimisation algorithm. Hence, first, we choose cluster centers in the PSO algorithm to calculate cluster centers according to defined objective function; then, we will apply these centers to the Rough k-means clustering algorithm as the initial average. In applying the Rough K-means algorithm, we use the ratio of distances instead of the distance difference, as suggested in the following equation. Using distance ratio had produced a better result for outliers clustering. Additionally, we have used a single-object particle swarm algorithm to determine w1 parameters and ε threshold values.

$$\frac{\mathrm{d}\ \mathrm{x_n}, \mathrm{m_i})}{\mathrm{d}\left(\mathrm{x_n}, \mathrm{m_j}\right)} \leq \xi, \qquad \mathrm{d}\ \mathrm{x_n}, \mathrm{m_i}) - \mathrm{d}\left(\mathrm{x_n}, \mathrm{m_j}\right) \leq \xi. \tag{1}$$

3.4 | Predicting Status of Some Lines and Stations

We use the delayed equation and function of time series. We can utilize linear models like ARIMAX and nonlinear models like ANNs to find this function. In this paper, we have used GDMH neural network to predict two-variable time series. Generally, to expect single-variable time series, we could say that representation of time series is a function of delayed samples, and our purpose is to find this function. In other words, with this assumption that:

$$\mathrm{t}) = \mathrm{f}\left(\mathrm{x}\ \mathrm{t} - \mathrm{d}_1), \mathrm{x}\ \mathrm{t} - \mathrm{d}_2), \dots, \mathrm{x}\ \mathrm{t} - \mathrm{d}_n)\right), \qquad \mathrm{Delay} = \{\mathrm{d}_1, \mathrm{d}_2, \dots, \mathrm{d}_n\}. \tag{2}$$

The main goal is obtaining function f, which describes time series in the best possible way. In this paper, we have used GMDH neural network to predict two-variable time series. In this type of neural network, suppose x is input and y is the model's output; the relation between input and output is mentioned below:

$$\mathrm{y} = \mathrm{a}_0 + \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{a_i}\ \mathrm{f_i}. \tag{3}$$

In this stage, we test the trained network and then evaluate obtained results to predict lines status.

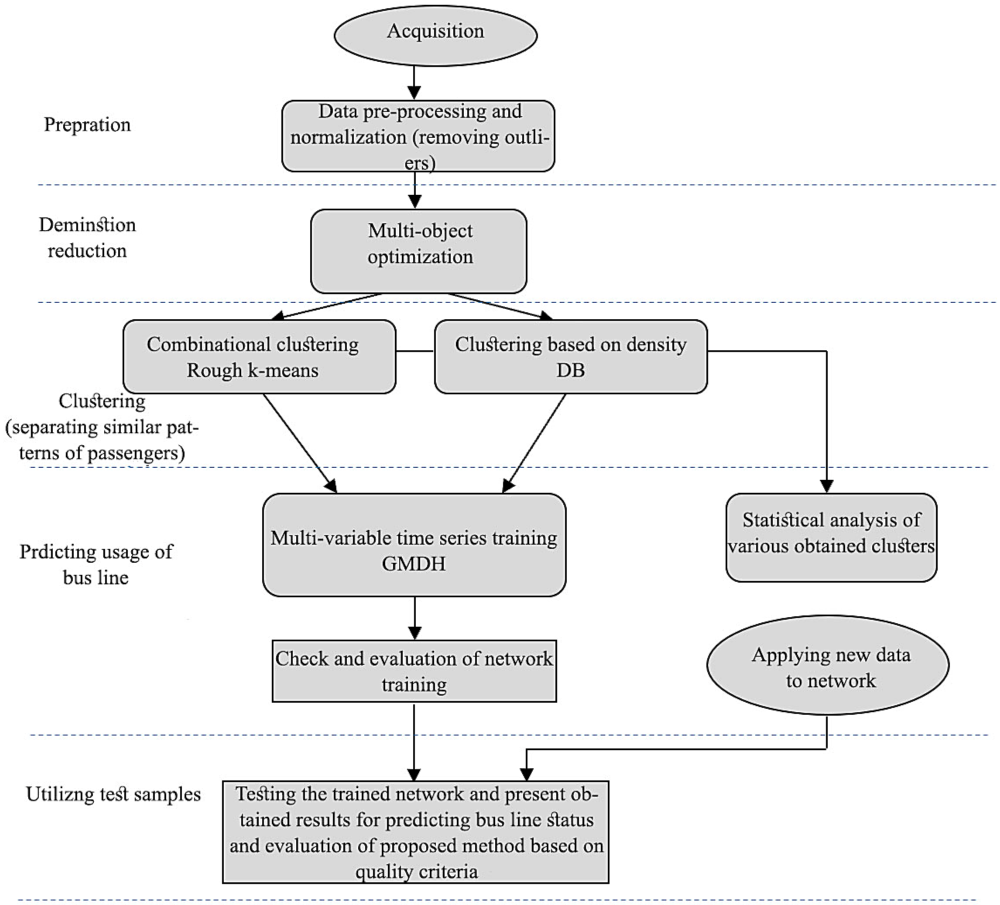

Fig. 1. A general framework for evaluation of behavioral patterns of the passenger.

This study has been performed utilizing smart card data and collected information from the automatic transportation system of Isfahan Card. The obtained data is related to lines (1) to (3) of express bus and lines (9), (21), (86), (91), and (4) of the urban bus, which all are in the Isfahan metropolis. The total number of stations correspond to these eight lines is 151 stations.

The given range for datasets is one month (October 2017) and includes (6, 571, 123) samples that have been acquired from 880,000 smart cards. Each of traffics has an identification number for passengers in a time interval of traffic and name and specifications of traffic station and name and specification of destination and information of the vehicle. Additionally, in case of traffic with the bus (internal or external), the direction of the trip is mentioned. AFC requires that passengers have extended their smart cards. To perform batch analysis of passengers, first, we should reconstruct the trip chain according to performed traffics with a smart card. If we use raw traffics to describe trip behavior, passengers involved in multi-stage trips will be considered more active than single sage passengers, which bias the classification process.

Trip chains are reconstructed by a two-stage method. The first stage is based on the following assumptions:

I. Assumption of the closest station: the passenger will place in a station nearest to his living location for specific traffic, and the beginning of the trip is from there.
II. Daily symmetry assumption: the last station is on symmetry to begin station or is close to it. In each traffic, we will predict the distance to all available stations in a line, and the nearest station is considered as the destination of the current trip. If the selected destination and following locations have a reasonable distance for walking, a location close to the current location is considered as the next station; otherwise, destination recognition will not be performed.

4 | Simulation Results

This study has been performed utilizing smart card data and collected information from the automatic transportation system of the Isfahan Card service. As it is mentioned, the user data are related to 8 bus lines with 985,448 samples (after excluding 11% of trips with low transactions) from 151 bus stations in Isfahan city. Additionally, these data have seven feature which is mentioned below:

I. Card Identification number (Card-ID).
II. Bus Identification number (Bus-ID).
III. Transaction Type (Trn-Type).
IV. Charge numbers (Ch-num).
V. Charge amount (Ch-amount).
VI. Cord number (card-num).
VII. Bus Line number (Bus-Line).

4.1 | Applying Feature Selection Algorithm

In the first step of applying data to the feature selection algorithm, the main features and the number of clusters between 2 to 15 are involved in the form of 4-bit gray code. Thus $n_f = 7$ and $k_c = 4$, and each particle will be in an 11-dimensions particle swar algorithm. Use parameters in the BMOPSO optimisation algorithm are mentioned in *Table 3*.

Table 3. Parameters in the BMOPSO optimisation algorithm.

γ	Mu	C1	C2	Population	W	α	β	Iteration
2	0.1	1	2	100	0.45	0.1	2	10

Fig. 2 shows the estimated Pareto front and other members after five epochs of the feature section algorithm with primary data. Dominated members of the population who form the Pareto front are highlighted with red color. Aside from overwhelmed members, several selected features are mentioned.

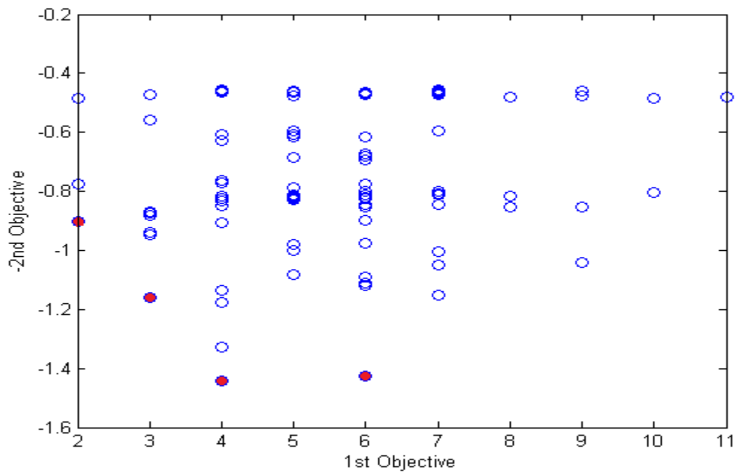

Fig. 2. Estimated Pareto front along with other members after five epochs of the feature selection algorithm.

As evident from the figure, minimizing two objective functions of several features and free space statistics validation criteria shows that price and density ratio features are the most significant features in the superstability of transportation smart card data. Finally, according to simulated results in ranking feature selection, we could consider the features as below:

I. Card-ID.
II. Bus-ID.
III. Bus-Line.

Daneshvar et al. | J. Appl. Res. Ind. Eng. 10(1) (2023) 11-24

IV. Ch-num.
V. Card-num.
VI. Trn-Type.
VII. Ch-amount.

Also obtained results for feature numbers of 4 to 7 with cluster number 1/100 show an optimized number of clusters is 9. According to the obtained Pareto front, we conclude that only 4 of the most influencing features in clustering criteria and 9 clusters will not have significant error. In the clustering layer, we have used only these four features.

4.2 | Data Clustering

Used parameters for the PSO algorithm are mentioned in *Table 4*. Also, paraments of Rough k-means algorithm are obtained from imperialist competitive optimisation algorithm with cost function of Davis Bolden clustering criteria. *Fig. 3* used parameters in the PSO algorithm.

Table 4. Parameters in the PSO algorithm.

Iteration (MAX)	Population	# of Empires	A	B	ζ	Mu
1000	100	15	1	2	0.1	0.1

Fig. 3 shows the training curve for the proposed algorithm and PSO-Rough k-means algorithm. The proposed method outperforms the PSO method, at least in cost function minimization.

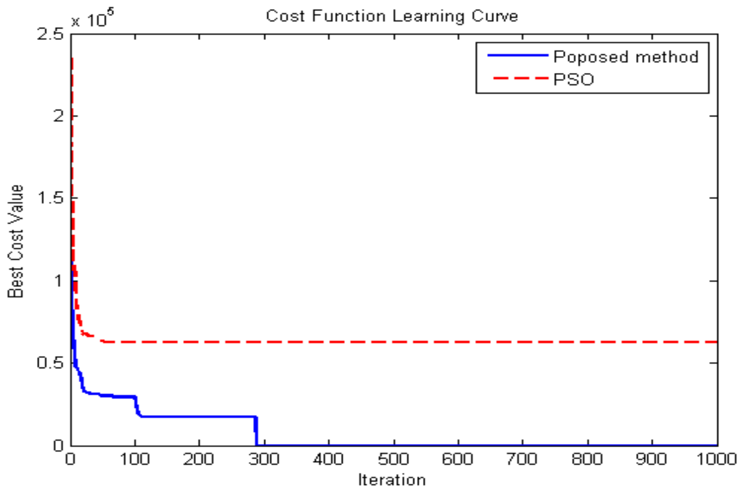

Fig. 3. Training curve of the proposed method and PSO-Rough k-means.

First, obtained clusters are studied according to the performed transactions' number in each station during days of the week (except Friday, excluded data in this analysis.)

Evaluating the activity of these stations shows that there are two general categories in clusters (*Fig. 4*). The first category includes clusters in which the number of daily transactions has had some peaks and are usually used during all hours of the day. Six clusters from 9 clusters have these properties, including 75.5% of stations (114 stations). The second category comprises clusters in which the amount of transactions in corresponding statins in a half-day is different from another half-day. These states include clusters with 24.5% (37 stations). And we could say that users use these stations to their workplace in the center of the city. *Fig. 4* shows a state of stations with usage peak in hours 7 to 8 in the morning.

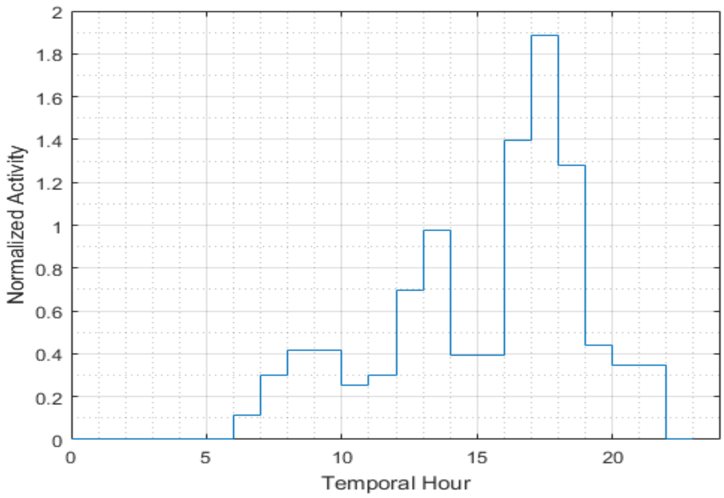

Fig. 4. Plots of two clusters that had usage peak in the morning.

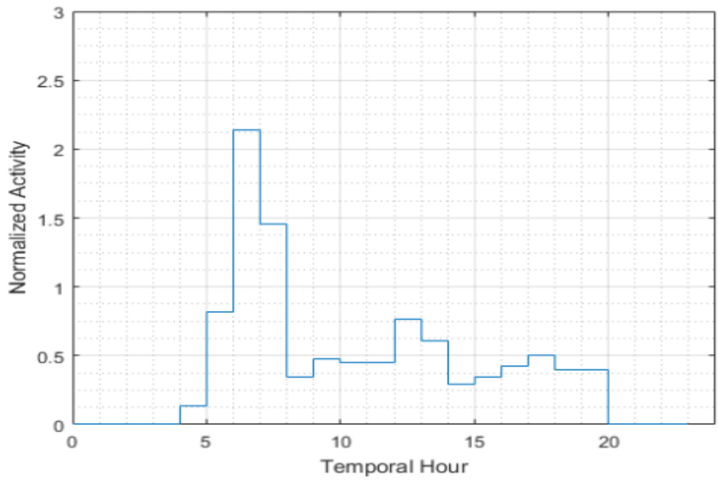

Fig. 5. Cluster plot which shows usage peak in post morning.

Fig. 5 shows the reverse state of two previous clusters which usage peak is in the evening in hours 17 to 18. A number of these stations are industrial areas, and many of them are in the city center. By studying this cluster, we notice that these stations correspond to individuals who have used these statins way back from industrial or official centers toward their homes. No significant usage from these stations is recorded for the rest of the day.

Fig. 6 shows the distribution of all clusters on days of the week. The aim is to evaluate passengers' behaviors during days of the week. As it is evident from this figure, in business days (Saturday to Wednesday), most of the usage was related to individuals if fifth, seventh, eighth, and ninth clusters, and in weekends (Thursday and Friday), most uses were related to first and third and sixth clusters. According to the plot, we could say that fifth, seventh, eighth, and ninth clutters related to users who most of them are a student it is the transportation system for going (or returning) from the workplace. On the other hand, first, third, and sixth clusters include users who have used public transportation to spend their weekends in most visited areas or found to the suburbs of Isfahan (by going to available terminals in the city) on their weekends. The second and fourth clusters have almost similar distribution on both business days and weekends. These clustered include industrial areas users who also go to their workplace as shifts in the last days of weeks.

Daneshvar et al. | J. Appl. Res. Ind. Eng. 10(1) (2023) 11-24

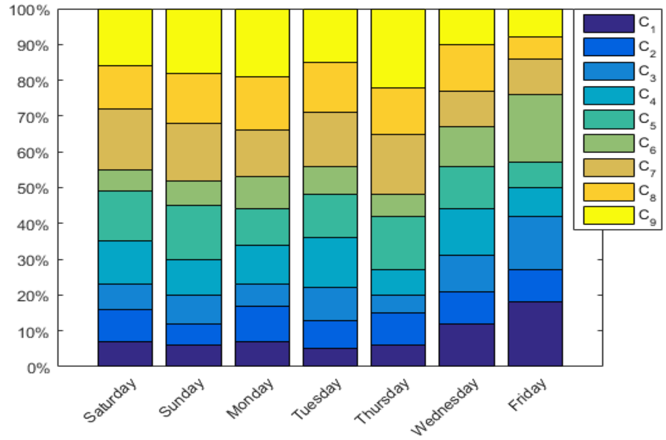

Fig. 6. Usage percent plot of each cluster during days of the week.

4.3 | Station Prediction

In this section, we have discussed the prediction of two different bus stations in terms of clustering. The number of prepared temporal samples for a week in each station is 2016. Delay vectors have been considered as below, and prediction of one day earlier is desired. Corresponding delays have been considered for a half-hour, one hour, one day, and two days.

Delays = [6 12 288 576].

Time series related to specified station from line 91 is shown in *Fig. 7*. Moreover, *Fig. 8* has shown the actual time series and estimated time series for training and test data station correspond to line 91, respectively. We should note that 80% of data is used for training, and 20% is used for testing.

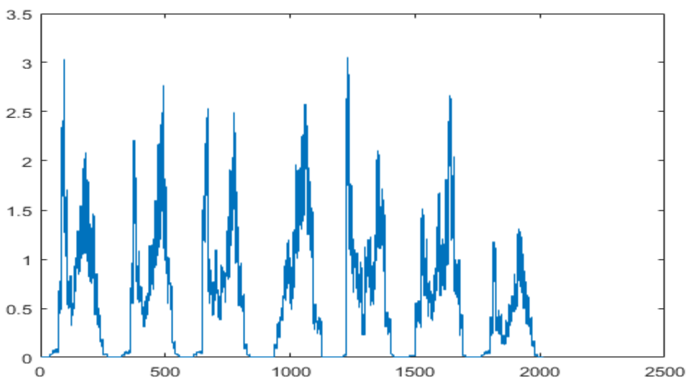

Fig. 7. Time series related to specified stations from line 91.

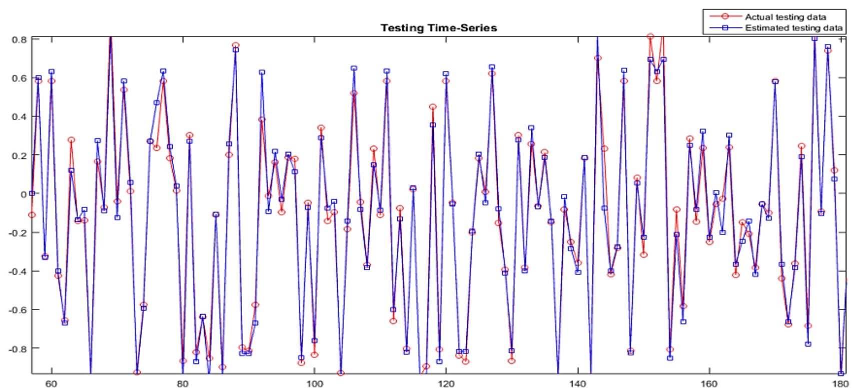

Fig. 8. Actual and estimate time series for test data in station related to line 91.

Fig. 9 has shown regression for training data and test data, respectively. As it is evident from these figures, prediction performance in this state is significantly more robust.

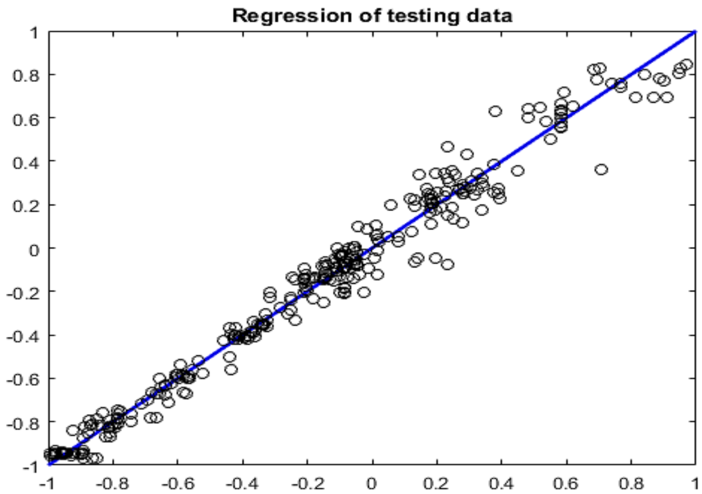

Fig. 9. Regression for test data.

5 | Conclusion

In this paper, we analyzed obtained data from the smart card of express transportation systems using data mining and artificial intelligence techniques. Three main axes considered in this analysis include:

I. Clustering stations and evaluating their activity statues.
II. Clustering from users (passengers) pint of view and comment usage behavior of passengers during the week.
III. Predicting status of two different stations using two-variable time series.

Obtained results for data related to one month included in 9 clusters. Stations have been placed in three general statuses: stations with high transactions in the morning, stations with the high transaction in the post morning, and stations with transactions during the day. This category has devoted most stations to itself. Results have shown that price and density ratio features are the most influencing features inseparability of transportation smart card data. According to obtained Pareto front, it has shown that the use of four influencing features includes card identification number, bus identification number, bus line number, and charge number in clustering criteria, will not generate a significant error and in the clustering layer, we have only used from these features. In the passenger's behavior section, the percentage of each cluster during days of the week is presented. The first category was clustered a whole number of transactions during the day had some peaks and generally, these stations have been used during all hours of a day. This category includes 6 clusters which form 75.5% of stations (114 stations). The second category includes clusters whose transactions in the different stations in half of the day were different from another half of day, which cost 24% station (37 stations). In the last step, the status of two distinct stations in form of time series is considered and the prediction of one day ahead has been predicted using a neural network. Behavior extraction of old stations is an important step for the planning of managers for different stations in terms of usage amount, which could be utilized in infrastructure plans. Also, we could utilize analysis of passenger's behavior and identification and separation of stations in terms of time which passenger traffic volume is high, to planning transportation systems. Finally, appropriate moving of buses and appropriate scheduling in stations which have passenger traffic load, lead to improve statue and reduce passenger volume.

For future research, we could use feature combination methods like principal component analysis PCA or independent component analysis in feature selection. The data analysis section could use required preprocessing processes like eliminating outliers and filling lost features before applying them to the main algorithm. Also, we could increase the number of objective functions.

Daneshvar et al. | J. Appl. Res. Ind. Eng. 10(1) (2023) 11-24

References

[1] Soltani, A. (2012). Public transportation in response to a common challenge in developing countries. *Quarterly report of Fars province building engineering organization*, *16*(55), 1-71. https://www.fceo.ir/NewsShow.Aspx?NewsId=Boxn5Q5U4vUN3ZVhMBrxgA

[2] Poorahmad, A., Omranzadeh, B. (2012). Evaluating and presenting strategies for developing the BRT transportation system in the metropolis of Tehran using the SWOT model. *Quarterly journal of urban research and planning*, *3*(11), 17-36. (**In Persian**). https://www.sid.ir/paper/220118/fa

[3] Guo, J. Y., & Bhat, C. R. (2001). *Representation and analysis plan and data needs analysis for the activity-travel system* (No. 0-4080-1). University of Texas at Austin. Center for Transportation Research. https://rosap.ntl.bts.gov/view/dot/4865

[4] Jabbari, E., & Saeedi, S. (2018). Solving the public transportation problem using linear programming and hybrid PS-ACO meta-heuristic method. *Journal of transportation engineering, 11*(2), 353-371. (**In Persian**). https://www.magiran.com/paper/2072392?lang=en

[5] Momeni, E., & Darvishzadeh, M. M. (2017). Investigating the nature and dimensions of passenger lines and the feasibility of using high-speed bus lines simultaneously with it in Tehran. *Rahvar research studies*, *6*(23), 43-72. (**In Persian**). https://ecc.isc.ac/showJournal/25525/101776/1316268

[6] Tavakoli Kashani, A., Mansoori Kargar, H. (2015). The effect of bus Drivers behavior of Tehran Branch Company on urban accidents. *Rahvar research studies, 5*(16), 39-62. (**In Persian**). https://www.noormags.ir

[7] Ahmadvand, A. M., Mohammadiani, Z., and Khodadadi, H. (2015). Dynamic modeling of sustainable urban transportation system in order to improve and improve traffic. *Management research in Iran*, *19*(2), 52-31. (**In Persian**). http://ensani.ir/fa/article/download/345444

[8] Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: a literature review. *Transportation research part c: emerging technologies*, *19*(4), 557-568.

[9] Agard, B., Partovi Nia, V., & Trépanier, M. (2013). Assessing public transport travel behaviour from smart card data with advanced data mining techniques. *World conference on transport research* (Vol. 13, pp. 15-18). http://datawisdom.ca/paper/2013-Transport.pdf

[10] Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings volumes*, *39*(3), 399-404.

[11] Dzikrullah, F., Setiawan, N. A., & Sulistyo, S. (2016). Implementation of scalable k-means++ clustering for passengers temporal pattern analysis in public transportation system (BRT Trans Jogja case study). *6th international annual engineering seminar (InAES)* (pp. 78-83). IEEE. https://doi.org/10.1109/INAES.2016.7821911

[12] Ma, X. L., Wang, Y. H., Chen, F., & Liu, J. F. (2012). Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang university science C*, *13*(10), 750-760. https://doi.org/10.1631/jzus.C12a0049

[13] Ma, X., & Wang, Y. (2014). Development of a data-driven platform for transit performance measures using smart card and GPS data. *Journal of transportation engineering*, *140*(12), 04014063. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000714

[14] Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., & Varshavsky, A. (2010). A tale of two cities. *Proceedings of the eleventh workshop on mobile computing systems & applications* (pp. 19-24). https://doi.org/10.1145/1734583.1734589

[15] Alfred Chu, K. K., Chapleau, R., & Trepanier, M. (2009). Driver-assisted bus interview: Passive transit travel survey with smart card automatic fare collection system and applications. *Transportation research record*, *2105*(1), 1-10.

[16] Park, J. Y., Kim, D. J., & Lim, Y. (2008). Use of smart card data to define public transit use in Seoul, South Korea. *Journal of the transportation research record*, *2063*(1), 3-9. https://doi.org/10.3141/2063-01

[17] Trépanier, M., & Morency, C. (2010). Assessing transit loyalty with smart card data. *12th world conference on transport research*. Lisbon. Portugal. https://www.researchgate.net

[18] Trépanier, M., Morency, C., & Agard, B. (2009). Calculation of transit performance measures using smartcard data. *Journal of public transportation*, *12*(1), 79-96.

[19] Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of intelligent transportation systems*, *11*(1), 1-14.

[20] Trépanier, M., & Vassiviere, F. (2008). Democratized smartcard data for transit operator. *15th world congress on intelligent transport systems and its America's 2008 annual meetingits americaerticoits japantranscore*. New York NY, United States.

[21] Ahn, J., & Han, R. (2013). Personalized behavior pattern recognition and unusual event detection for mobile users. *Mobile information systems*, *9*(2), 99-122. DOI: 10.3233/MIS-130155

[22] Yan, N., & Au, O. T. S. (2019). Online learning behavior analysis based on machine learning. *Asian association of open universities journal*, 14(2), 97-106

[23] Osoba, O., & Davis, P. K. (2019). An artificial intelligence/machine learning perspective on social simulation: New data and new challenges. *Social-behavioral modeling for complex systems*, 443-476. https://doi.org/10.1002/9781119485001.ch19

[24] Ceja, E. G. (2021). *Behavior analysis with machine learning using R*. Chapman and Hall/CRC.

[25] Aliahmadi, A., Jafari-Eskandari, M., Mozafari, A., & Nozari, H. (2016). Comparing linear regression and artificial neural networks to forecast total productivity growth in Iran. *International journal of information, business and management*, *8*(1), 93-113.

[26] Udoka, C. G. (2020). The impact of passenger's traffic on exchange rate and economic growth in Nigerian aviation industry. *International journal of research in industrial engineering*, *9*(4), 364-378.

[27] Habibi, R. (2021). Poisson regression model with change points. *Journal of applied research on industrial engineering*, *8*(Special Issue), 1-8. https://doi.org/10.22105/jarie.2021.229051.1149

[28] Coello, C. A. C., Lamont, G. B., & Van Veldhuizen, D. A. (2007). *Evolutionary algorithms for solving multi-objective problems* (Vol. 5, pp. 79-104). New York: Springer. https://link.springer.com/book/10.1007/978-0-387-36797-2

Daneshvar et al.| J. Appl. Res. Ind. Eng. 10(1) (2023) 11-24