Scientia Iranica A (2019) 26(5), 2689–2702

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
http://scientiairanica.sharif.edu

Airline delay prediction by machine learning algorithms

H. Khaksar and A. Sheikholeslami*

Department of Transportation Engineering and Planning, School of Civil Engineering, Iran University of Science & Technology, Tehran, Iran.

Received 24 May 2017; received in revised form 15 October 2017; accepted 23 December 2017

KEYWORDS
Flight delay predictor;
Airline delay;
Data mining;
Machine learning algorithms;
Visibility distance.

Abstract. Flight planning, as one of the challenging issues in the industrial world, is faced with many uncertain conditions. One such condition is delay occurrence, which stems from various factors and imposes considerable costs on airlines, operators, and travelers. With these considerations in mind, we implemented flight delay prediction through the proposed approaches that were based on machine learning algorithms. The parameters that enabled effective estimation of delay were identified and then, Bayesian modeling, decision tree, cluster classification, random forest, and hybrid method were applied to estimate the occurrences and magnitude of delay in a network. These methods were tested on a US flight dataset and then, refined for a large Iranian airline network. Results showed that the parameters affecting delay in US networks were visibility, wind, and departure time, whereas those affecting delay in the Iranian airline flights were fleet age and aircraft type. The proposed approaches exhibited an accuracy of more than 70% in calculating delay occurrence and magnitude for both the US and Iranian networks. It is hoped that the techniques put forward in this work will enable airline companies to accurately predict delays, improve flight planning, and prevent delay propagation.

1. Introduction

Over the past century, airlines have grown from simple contract mail carriers to intellectually appealing businesses. However, the airline industry is very competitive, dynamic, and random in nature, thereby giving rise to uncertainties. One such uncertainty is flight delay, which is attributed to various factors, such as bad weather conditions, physical flaws, delayed flight arrivals, and crew related issues [1]. It is obvious that flight delays have a direct impact on passenger satisfaction and thus, on the economic performance of airlines. As estimated by the Total Delay Impact Study, the total cost arising from all US air transportation delays in 2007 was US.$32.9 billion [2]. Given this figure, airline companies have been concerned primarily with understanding and mitigating delays. Many researchers have also attempted to predict these uncertainties and ensure more reliable and robust flights. Achieving these goals requires evolution in approaches to airline planning, a development that is possible through a consideration of operational issues in the planning process.

Flight planning is a complicated process that involves many variables. As one of the tasks in flight planning, scheduling is performed using a step-by-step process that covers two separate phases, namely, the planning and operational phases. In each phase, certain sub-problems of low complexity occur and are solved sequentially [3]. The planning phase entails four steps, namely, flight scheduling, fleet assignment, fleet routing, and crew assignment. Each of these steps is implemented separately and used as the input to the next step. The operational phase involves

*. *Corresponding author. Fax: +98 21 88331528*
E-mail addresses: hasankhaksar@yahoo.com (H. Khaksar); sheikh@iust.ac.ir (A. Sheikholeslami)

doi: 10.24200/sci.2017.20020

revenue management, gate assignment, and irregular operation steps. Revenue management is one of the most important components of the operational phase and is used by airlines as the basis for controlling the number of seats sold at different fares. Gate assignment revolves around the allocation of gates to boarding and departure. When an airplane descends to a height at which it is cleared for landing, the control tower assigns gates to the airplane on the basis of the activities of other airplanes. The last step is irregular operation, which concerns the tendency of flight program implementation to deviate from plans. When a flight delay occurs, the airline operation control center executes activities designed to prevent delay propagation to other flights.

An effective flight planning process is ensured through three approaches: the classic approach, the robust approach, and disruption management. In the classic approach, the problems that occur in each step of the planning phase are separately implemented or the problems in a number of steps are integrated and solved simultaneously. Details on this approach can be found in [4-6]. In the robust approach, the problems occurring in one or more steps of the planning phase are solved in relation to the operational phase. More information on this method is provided in [7,8]. Disruption management involves the consideration of the operational problems that occur during flight planning. As indicated by the term, this approach is used when a disruption occurs [9,10].

The Bureau of Transportation Statistics of the US Department of Transportation releases national-level information about the on-time arrival performance of flights. Such information on all US airlines for a period of 12 years is shown in Figure 1. As shown in this figure, delays can be caused by seven factors: aircraft delay, weather condition, national aviation system delay, security problems, late flight arrival, flight cancellation, and flight diversion [11].

Delays are generally classified into two types, namely, arrival and departure delays. A flight with less than 15 minutes delay with respect to arrival and departure times printed on the ticket is considered on time. The causes of airport delays include weather, airport congestion, luggage loading, and passenger connection issues. Correspondingly, airline delay prediction is a very complex pragmatic and scientific challenge. It becomes even more important when the robust approach and disruption management have already been applied. Robust planning is intended to decrease delays and prevent delay propagation to other flights. Its first stage therefore involves determining delay. Disruption management should also be directed toward delay prediction.

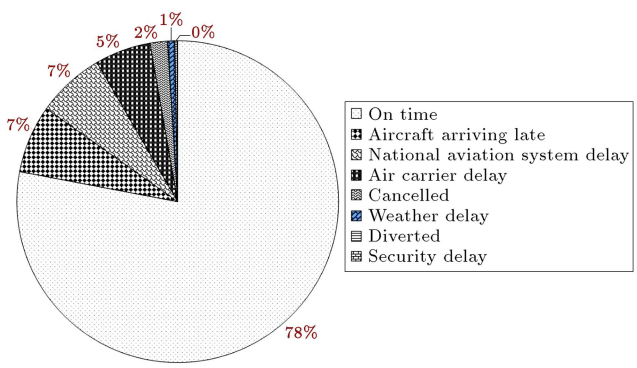

Figure 1. On-time arrival performance.

To accurately predict delays, improve flight planning, and prevent delay propagation, we put forward new delay prediction approaches that are based on data mining and machine learning. These methods are Bayesian modeling, decision tree, cluster classification, random forest, and hybrid algorithms. The approaches are applied in experiments to the US and Iranian flight networks. The rest of the paper is organized as follows. Section 2 presents an overview of the literature and Section 3 describes well-established Data Mining Techniques (DMTs) for airline delay prediction. Section 4 explains the proposed methods and Section 5 discusses the experimental datasets and results. Finally, Section 6 provides the conclusions and recommendations.

2. Literature review

Airline operations are highly complex processes that are intended to regulate many expensive, tightly constrained, and interdependent resources, such as the crew, aircraft, airports, and maintenance facilities. Many studies have been carried out on airline planning problems, but only a few have been performed on the characteristics of airline delays and the prediction of delay statistics. Delays occur when an event takes place later than the time at which it is planned, scheduled, or expected to happen [12]. Delays in departure can occur due to bad weather conditions, seasonal and holiday demands, airline policies, technical issues such as the problems in airport facilities, luggage handling and mechanical apparatus, and accumulation of delays from preceding flights (Figure 2). In a study conducted by Tu et al., distribution of departure delays at Denver International Airport was estimated with the help of a probabilistic model, which relied on a combination of genetic algorithms and expectation-maximization algorithm [13].

In a study carried out by Mueller and Chatterji, the probability density function of delays in major airports of America was estimated by normal and Poisson distributions [14]. After using the density functions to model the departure delay, enroute delay, and arrival delay, Poisson distribution was found to

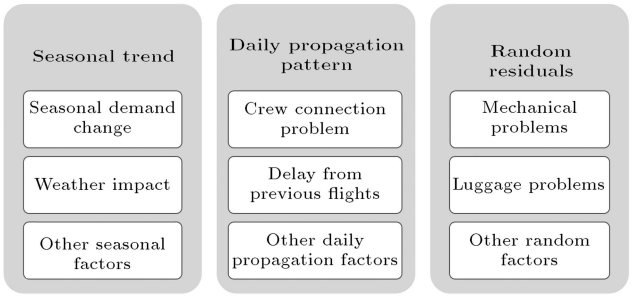

Figure 2. Factors influencing departure delays.

be capable of predicting departure delays, but it was outperformed by normal distribution in the prediction of enroute and arrival delays. Avijit et al. proposed an aggregate and strategic model for average delays and cancellation probabilities, and applied the model to the context of LaGuardia Airport [15]. In a research conducted by Sridhar et al., weather-related flight delays and cancellations at the national, regional, and airport levels were predicted by regression models and neural networks [16]. Lu presented Bayesian network and decision tree models to describe flight delays and stated that accurately predicting such delays was very difficult [17]. Lu et al. compared the performances of Naive Bayes, decision tree, and neural network algorithms in predicting delays on the basis of large datasets [18]. The authors observed that the decision tree algorithm had the best performance presenting a prediction confidence of 70% [18]. Bolaños and Murphy investigated the effects of airports on flight delays [19]. The researchers used a graph theory to analyze directed weighted graphs that represented the propagation of delays across the National Airspace System (NAS) of America. Their findings revealed that the New York area was responsible for 15% of injected delays and 9% of propagated delays in the NAS. Rebollo and Balakrishnan used random forests and regression models to predict air traffic delays [20]. The authors examined their models using the 100 most frequently delayed links in the NAS. They found that the average test error over the 100 links was 19% and the average median test error over the links was 21 minutes. Oza et al. predicted flight delays on the basis of certain data patterns observed from previously acquired flight information [21]. The researchers used classifier models to predict airline delays and achieved an estimation accuracy of 64.08% with the OneR algorithm.

Allan et al. probed into delays at New York city airports for two years to determine the major causes of delays and discovered that delays were caused primarily by the weather inside and outside the terminal areas and high winds [22]. In a study carried out by Wu, the effects of short buffer times and random operation disruptions on the consequent flight delays were investigated [23]. This study demonstrated the necessity of incorporating the stochasticity of daily operations into airline schedules. The author asserted that schedules could become robust and reliable only if buffer times were appropriately embedded and designed during scheduling. Wang et al. used queuing models to analyze how the responses to propagated delays varied by airport [24]. The same method was also used by Janic for the quantification of economic effects of flight delays [25]. Hsiao and Hansen adopted a statistical model to examine the effects of different parameters, such as time of day, congestion, and weather conditions on flight delays [26]. In a research conducted by Rupp, the primary causes of delay were explored by studying important factors such as weather condition, station type (hub versus spoke), and seasonality [27].

In an article by Boswell and Evans a probabilistic mass function was used to classify the delays and then, delay build-up for the subsequent flights was analyzed with the help of a transition matrix [28]. After the analysis of cancellations, the conditional probability of flight cancellation in the presence of delay build-up from preceding flights was determined. In a study carried out by Chen et al., departure and arrival delays were successfully predicted with the assistance of a model developed based on a fuzzy Support Vector Machine (SVM) with a weighted margin [29]. The evaluations carried out based on the definition of five grades of delays demonstrated the high prediction accuracy of the ensemble fuzzy support vector machine with a weighted margin as compared to standalone SVM.

The studies reviewed are summarized in Table 1. As can be seen, although several studies have been devoted to predicting airline delays, such prediction remains a challenge to flight planners. Many parameters are involved in the occurrence of delays; thus, delay forecasting necessitates the assessment of large databases.

3. Data mining for airline delay prediction

The analysis and design of complicated and large-scale systems with many variables require new methods that can identify, classify, and analyze voluminous data. Accordingly, researchers put forward data mining approaches to identifying, collecting, classifying, and ranking or generating and storing valuable information from such databases. These approaches learn knowledge from an available databank with the help of machine learning algorithms. Data mining can also be used to predict future events.

DMTs are a branch of artificial intelligence that have been used since 1960 [30]. In such techniques, prediction is a decision problem with the following components [31]:

Table 1. Summary of literature review.

Author	Methodology	Results
Tu et al. [13]	Probabilistic models	The probability of delay can be predicted when delays are less than 2 hours.
Mueller et al. [14]	Normal and Poisson distributions	Poisson distribution is better for departure delays and arrival delays can fit the normal distribution.
Lu [17]	Bayesian network and decision tree	It is very difficult to accurately predict flight delays.
Lu et al. [18]	Naïve Bayes, decision tree, neural network	Decision tree has the best performance presenting a prediction confidence of 70%.
Oza et al. [21]	OneR algorithm	Delays are predicted with 64.08% of accuracy.
Chen et al. [29]	Fuzzy support vector machine	Fuzzy support vector machine with weighted margin is more accurate than a standalone support vector machine.

- The state of the world, which indicates the available decisions, actions, or answers to an inference problem occurring in the real world;
- An answer $j \in J$ to a problem in a prediction task, which is the prediction of a future observation;
- A utility function, which defines a reward for predicting an unknown future observation;
- A specification of a current belief about the state of the world, which is described by a posterior predictive distribution.

The basic objective of these procedures is to detect the correlation between parameters and predictions of the future of a system. Machine learning is typically used to predict the changes in systems that perform tasks associated with artificial intelligence. Such tasks involve recognition, diagnosis, planning, and prediction, among other activities.

The literature presents various types of DMTs, which can be categorized into two groups [32]: Classification techniques and cluster analysis techniques. Classification methods include decision tree, Bayesian classification, rule-based classification, support vector machine, and lazy learners, whereas cluster analysis methods encompass clustering and partitioning methods, hierarchical methods, density-based methods, and grid-based approaches. From another point of view, DMTs can be classified into supervised (predictive or directed) and unsupervised (descriptive or undirected) methods. Classification is a supervised data mining method, whereas clustering is an unsupervised form of DMT. Both categories encompass functions that can identify different hidden patterns in large datasets.

In this work, we apply five methods of predicting airline delays, namely decision tree, random forests, Bayesian classification, a clustering method of the K-means algorithm, and a hybrid method (decision tree combined with clustering approach).

3.1. Decision tree

The typical structure of a decision tree consists of a root node, multiple branches, and a great number of leaf nodes. In this approach, the user develops a decision tree incrementally by partitioning the dataset into increasingly smaller subsets until reaching a tree with decision nodes and leaf nodes. In this tree, each internal node corresponds to a test on an attribute, each branch corresponds to a test result, and each leaf node represents a class label. The topmost node in the tree is the root node.

A typical decision tree that concerns the airline flight delay is shown in Figure 3. As indicated in the figure, a scheduled departure is the root and some other parameters, such as fleet age and visibility distance, are the leaves of the tree. The average age of the fleet is about 25 years. Although this age is older than the world standard, it is the average age reflected in our database, because of sanctions in previous years.

3.2. Random forest

Random forest is a concept falling under the general technique of random decision. This algorithm operates by creating a group of decision trees at training time and outputting the class that represents the mode of

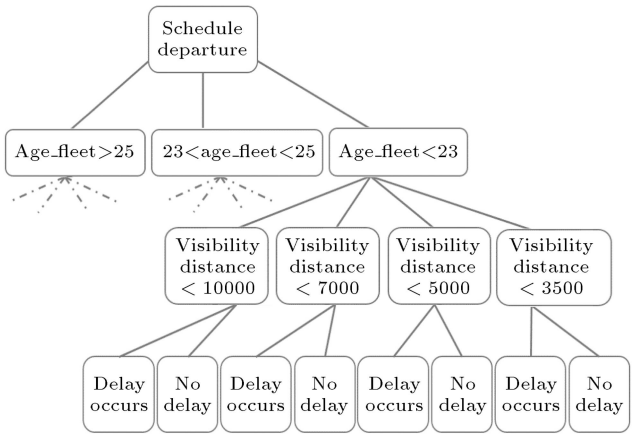

Figure 3. A typical decision tree for delay prediction.

classes or the mean prediction of the individual trees. Individual decision trees are generated using a random selection of attributes at each node to determine split. During classification, each tree casts a vote and the most popular class is returned. Using the random forests, the variance can be reduced by averaging the deep decision trees trained with different parts of the training set. To form random forests, tree predictors should be integrated in a way that each tree is dependent on the values of a random vector sampled independently and uniformly from all trees in the forest [33]. We use this approach to predict flight delays in our database. Rebollo and Balakrishnan used a random forest algorithm to predict departure delays of 2 to 24 hours in the NAS. Their results confirmed that random forests were suitable for predicting airline delays [20].

3.3. Bayesian classification

Being an offspring of Bayes's theorem, Bayesian classification has shown desirable accuracy and speed in handling large databases [32]. Bayesian classification is directed toward finding the best j in a supposed space (J) across a given training dataset X (data on flight delays). Accomplishing this necessitates identifying an assumption with the highest probability in dataset X. The term $p(j)$ is the prior probability or a priori probability of J (delay occurrence). $p(X)$ is the prior probability of X and $p(X|j)$ is the posterior probability of X conditioned on j.

Bayes' rule can be used to obtain expressions for $p(j|X)$ in terms of $p(X|j)$, which is assumed to be known (or estimable):

$$p(j|X) = \frac{p(X|j)p(j)}{p(X)}. \tag{1}$$

In learning scenarios, such as those featuring the prediction of flight delays, a learning algorithm considers the assumptions that underlie J and tries to find assumption j of J with the maximum probability of occurrence (or the assumption with the highest likelihood of occurrence). The Maximum A Posteriori (MAP) is an assumption with the maximum likelihood that is calculated using Bayes' rule and the posterior probability:

$$h_{\mathrm{MAP}} = \arg_{j\in J} P(j|X) = \arg_{j\in J} \max \frac{P(X|j)P(j)}{P(X)}$$

$$= \arg_{j\in J} P(X|j)P(j). \tag{2}$$

The class of the newly derived sample is then predicted by Bayesian classification. Thus, the class with the highest likelihood of occurrence (or V_{MAP}) is computed by identifying $a_1, a_2, ..., a_n$:

$$v_{\mathrm{MAP}} = \arg_{v_j\in V} \max P(v_j|a_1, a_2, \cdots, a_n). \tag{3}$$

In Eq. (3), v_{MAP} is the value with the maximum likelihood of occurrence and $a_1, a_2, ..., a_n$ are the values of each attribute.

Rewriting the formulation in Bayes' rule yields the following:

$$v_{\mathrm{MAP}} = \arg_{v_j\in V} \max \frac{P(a_1, a_2, \cdots, a_n|v_j)P(v_j)}{p(a_1, a_2, \cdots, a_n)}$$

$$= \arg_{v_j\in V} \max P(a_1, a_2, \cdots, a_n|v_j)P(v_j). \tag{4}$$

The occurrence probability of $a_1, a_2, ..., a_n$ is calculated by multiplying each of the classes.

$$v_{\mathrm{NB}} = \arg_{v_j\in V} \max P(v_j)\Pi_i p(a_i|v_j). \tag{5}$$

In this relationship, v_{NB} is the output of the Bayes classification. The number of $P(a_j|v_j)$ terms that are calculated in this procedure is equal to the number of classes multiplied by the number of outputs, which is less than the number of $P(a_1, a_2, \cdots, a_n|v_j)$ terms.

3.4. K-means clustering

Another approach used to predict flight delays is clustering, which comes in various types, such as K-medoids and K-means. We use the K-means method that is a type of partitioning approach because of its simplicity, rapid operation, and consequent applicability to large datasets. Nevertheless, this method has three limitations, namely handling empty clusters, measuring the sum of a square error higher than that encountered in the real world, and reducing the sum of a square with postprocessing. We resolve these shortcomings by using the methodology proposed by Singh et al. [34]. Some studies reported good results with the use of the K-means algorithm, evaluating it as one of the best clustering algorithms available [35]. In the algorithm, there is a set of d-dimensional real vectors of observations $(x_1, x_2, \cdots, x_n)$ (flight delays), in which c_j is the mean or weighted average of each

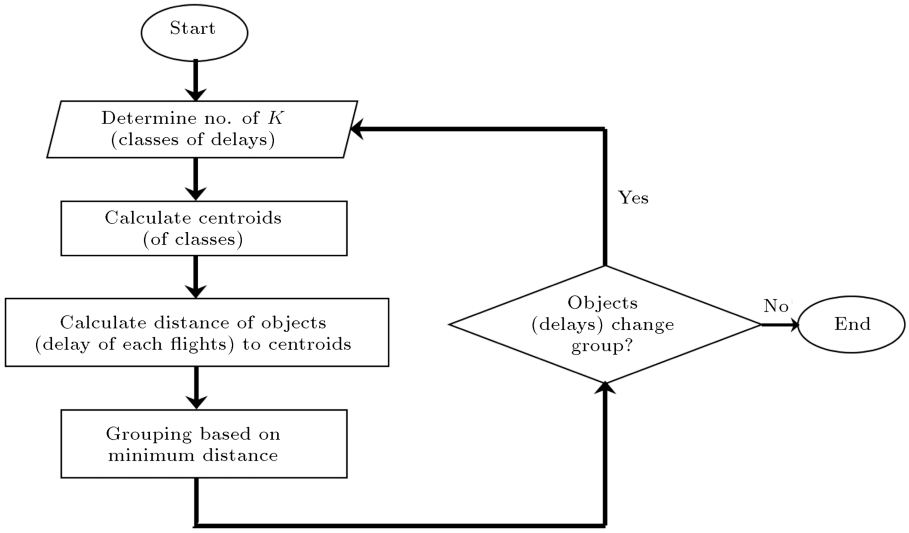

Figure 4. Process of K-means algorithm.

cluster. In the K-means algorithm, n observations will be divided into k sets in such a way as to minimize the Cluster Sum of Square Error (CSSE).

$$\min \ \text{CSSE} = \sum\nolimits_{j=1 \text{ to } k} \sum\nolimits_{x_i \in C_j} \|x_i - c_j\|^2. \qquad (6)$$

Figure 4 illustrates the process underlying the application of the K-means algorithm. First, k (the number of data classes) is determined on the basis of the nature of a database and the experience of a user. In this research, classes are set on the basis of the delays of each flight and the center of each cluster is calculated using the formulation below:

$$C_j^{t+1} = \frac{1}{|S_i|} \sum_{k, x_k \in S_i} x_k, \qquad (7)$$

where C_j^{t+1} denotes the new center of the cluster, $t+1$ represents a new class in the cluster, S_i is the ith set of the cluster, and x_k is the kth member of set i. Second, the distance of objects (delays) to each centroid is calculated and the objects are grouped in various clusters in accordance with their distance to the centers. The process is repeated until the members of the clusters differ from those in the previous iteration.

3.5. Hybrid approach

Clustering analysis is an important and commonly used data analysis technique. It can be used to describe a dataset and a decision tree can be applied to analyze the dataset. To improve the results of isolated approaches, we use a hybrid approach, which is a decision tree based on the clustering algorithm. In this method, clustering is adopted as a down-sampling preprocess for classification to reduce the size of a training set. The result is reduced dimensionality and a smaller, less complex classification problem that is easier and faster to solve. Thus, a decision tree that uses K-means clustering for classification is generated on the basis of the assumption that each cluster corresponds to a class. Such classification and assumption are adopted in this work. The process underlying the hybrid approach is shown in Figure 5. First, the usage and interaction of data are preprocessed. Then, an optional attribute selection process is applied to selecting only a group of attributes/variables or using all the available attributes/variables. Finally, a clustering algorithm is executed using training data. An essential requirement is to ensure that the number of generated clusters is the same as the number of class labels in a dataset. This

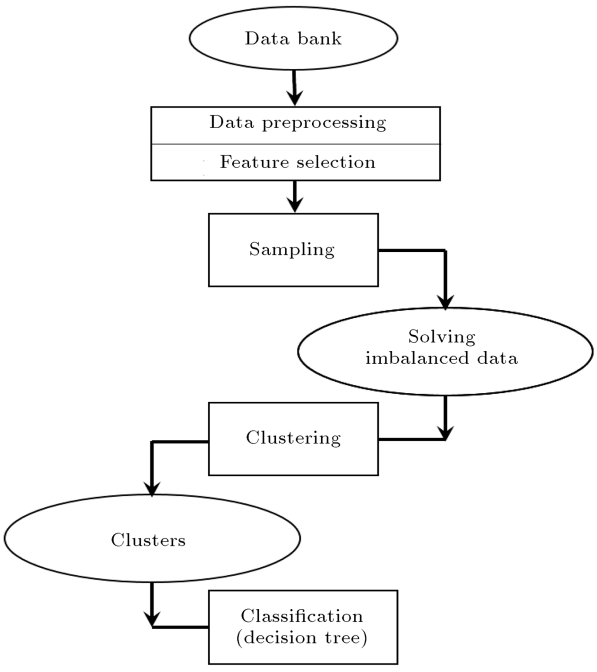

Figure 5. Process of hybrid approach.

identicality guarantees the derivation of a useful model that relates each cluster to one class.

4. Methodology

The proposed approaches were used to predict the flight delays of a large airline in Iran. Steps were taken to test the proposed approaches on the real-world data, which are presented in Figure 6. The first steps of the FDP conducted in this work were data preprocessing and feature selection. The flight dataset on the Iranian airline is very large; thus, analyzing it requires data preparation (data preprocessing is described in the next section). Another preprocessing task was the selection of effective variables from among various types of data, after which flight delays were predicted using decision tree, cluster, hybrid, random forest, and Bayesian classification. In the third step, the proposed approaches were evaluated and the best approach to FDP was determined through the comparison and verification of the approaches. This process was carried out two times; that is, delay occurrence and then delay magnitude were determined via FDP.

4.1. Data preprocessing

Airline datasets are huge and increasingly large in terms of the number of dimensions and instances. An essential step, therefore, is data preprocessing, which includes the tasks of data cleaning and data integration. We used an SQL data warehouse to manage our database. Table 2 describes the datasets and Table 3 shows the parameters of our model for the US and Iranian networks. Data on the US flight network and the weather conditions were obtained from [11] and [36]. On the basis of Table 3, the parameters that contribute to flight delays can be classified into seasonal trends and random residuals. Seasonal trends include seasonal demand changes, weather effects, and other seasonal factors. Seasonal demand changes encompass dates and days of the week. The meteorological conditions of flights were acquired from the Islamic Republic of Iran Meteorological Organization (IRIMO). Such conditions are related to visibility, which is a measure of the distance at which an object or light can be clearly distinguished. Other associated parameters include station pressure, temperature, and wind speed. Random residuals revolve around flight-related parameters, such as flight number, sector number, scheduled departure time, and aircraft type and age. Other factors associated with random residuals are origin and destination.

Table 2. Description of data used in the model.

Title	Description	
	Iran network	**The US network**
Time span	16 months	6 months
Number of airports	52	278
Number of aircrafts	35	—
Number of aicrcaft types	9	—
O-D pairs	96	5346
Number of operations	15428	2825647

4.2. Feature selection

A good understanding of feature selection is considerably advantageous in that it leads to improved model execution and enhanced understanding of the underlying structure and characteristics of the data.

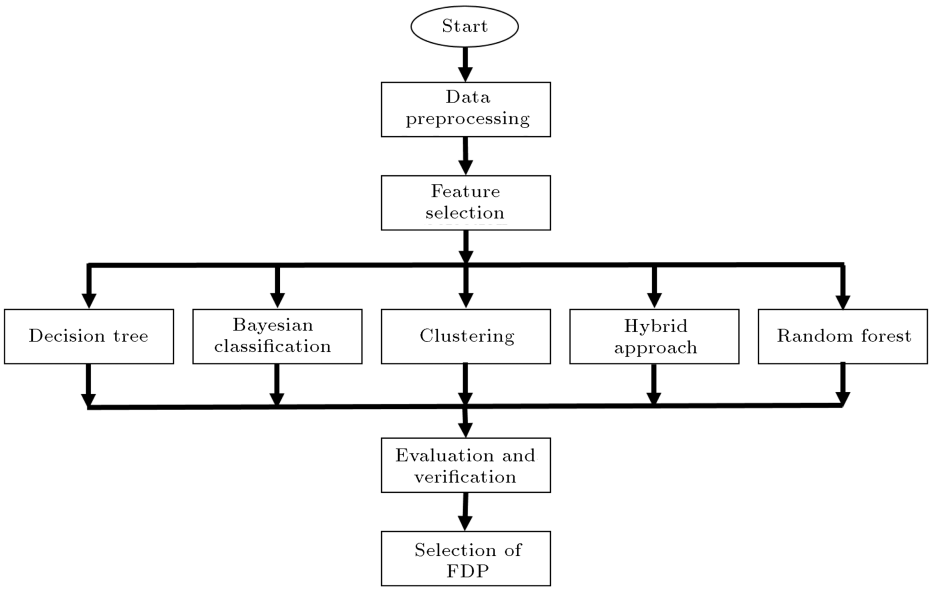

Figure 6. Process of FDP.

Table 3. Variables that are used in data mining.

Category of delay	The US network	Iran network
Seasonal trend	Visibility of origin and destination	Visibility of origin and destination
	Temperature of origin and destination	Temperature of origin and destination
	Station pressure of origin and destination	Station pressure of origin and destination
	Wind speed of origin and destination	Wind speed of origin and destination
	Date	Date
	Day of week	Day of week
Random residuals	Unique carrier	Flight number
	Origin city market ID	Number of sectors
	Scheduled departure time	Scheduled departure time
	Destination city market ID	Type of aircraft
		Age of aircraft
		Origin

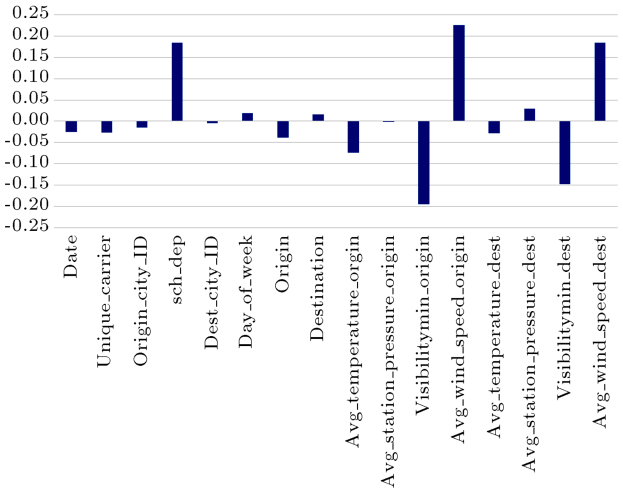

Figure 7. Effects of the variables on delay based on the correlation matrix (the US network).

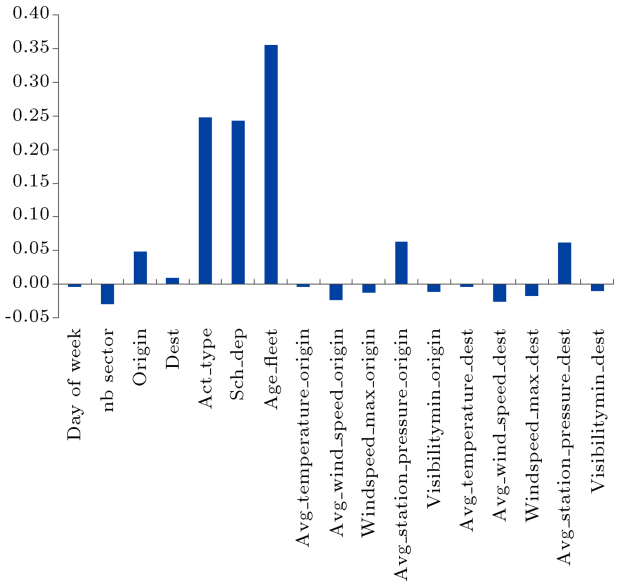

Figure 8. Effects of the variables on delay based on the correlation matrix (Iran network).

In this step, the correlation between each parameter and delays was determined using a correlation matrix. The outputs of the matrix are shown in Figures 7 and 8. Using a correlation matrix is a statistical technique that provides the correlations between multiple variables and consist of the coefficients of correlation between each variable and others. Put differently, it shows the relationship between two variables and the magnitude of this relationship. Matrix elements are correlation values that can vary from -1 to $+1$. A positive quantity shows a direct relationship between two variables, whereas a negative value for the correlation implies a negative or inverse association [37]. On the basis of the US network dataset, scheduled departure time, visibility of origin and destination, and wind speed at origin and destination exert the highest effects on flight delays in the US network (Figure 6). To accurately estimate flight delays, we used the aforementioned result to select the variables for use in our proposed methods. On the basis of the Iranian dataset, fleet age, scheduled departure time, aircraft type, and average station pressure strongly affect flight delays in the Iranian network (Figure 8). Weather conditions may be effective as parametric bases for predicting flight delays around the world, but parameters such as fleet specification and scheduled departure time are more effective for predictions in the Iranian context. The age of aircraft fleet in Iran is older than the standard, and congestion problems during peak hours may cause flight delays in Iranian airports. Therefore, the fleet age and scheduled departure time

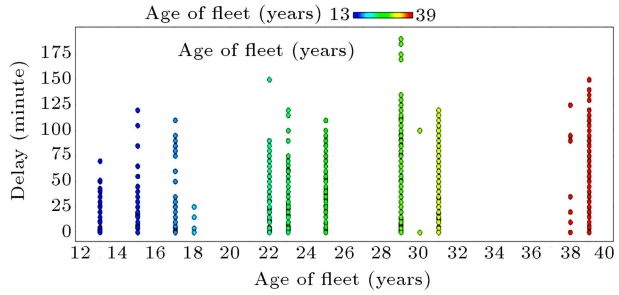

Figure 9. Effects of the age of the fleet on the flight delays.

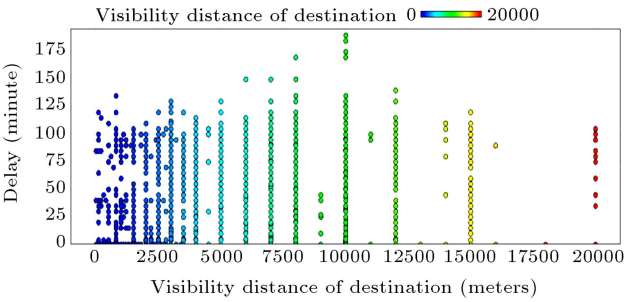

Figure 10. Effects of the visibility of destination on the flight delays.

are two effective parameters in flight delays in Iranian network.

The effects of fleet age on delays in the Iranian network are illustrated in Figure 9, which indicates that older fleets experience more delays than do younger fleets because of maintenance issue. The effects of visibility distance on delays in the Iranian network are presented in Figure 10, which shows that airports with low visibility distances experience more delays.

4.3. Improving the classification for imbalanced data

The databases treated in this work are imbalanced in nature, and records on delay magnitude are rare. The term "imbalanced dataset" refers to any dataset that contains a particularly rare class with significantly greater degree of importance than other classes. In an imbalanced dataset, the class with a higher number of instances is called a major class, whereas that having a relatively lower number of instances is referred to as a minor class. In our databases, flights with delays belong to the minor class, thus compelling us to use a sampling technique to solve this problem. Sampling is an effective approach to solving imbalanced data. Conventional classification methods perform poorly when a dataset is imbalanced, because such methods disregard class imbalance in their operations. Classic approaches treat data in the same manner, regardless of whether they belong to common or minority classes. They are aimed at optimizing overall accuracy without consideration for the relative distribution of each class [38]. Imbalanced data diminish the generalization of machine learning results. The specific sampling technique used in our work was Random Under-Sampling (RUS), which was one of the best techniques available [39]. In RUS, instances of a majority class are randomly discarded from a dataset, and rare instances are copied and distributed across the dataset. Thus, rare instances are intensified against common instances, resulting in slightly reduced total accuracy of results. Nevertheless, RUS results are reliable. Foregoing a sampling technique would mean more accurate data mining, but the accuracy of predicting the number of rare instances will be zero. In the airline industry, flights with delays are somewhat rare. Our application of RUS was intended to improve our results and achieve reliable prediction.

5. Experimental results

The proposed approaches were applied in actual airline networks, after which the results were compared to find the best solution. We used a machine learning analytics program called RapidMiner, which provided an integrated environment for data preprocessing, machine learning, deep learning, text mining, and predictive analytics [40]. We partitioned the data into testing (25%) and training (75%) sets.

5.1. Decision tree

We used the J48 pruned tree to estimate flight delays on the basis of the airline databases because of its good efficiency and accuracy [41]. J48 is a univariate decision tree, in which splitting is performed by one attribute at internal nodes. This method can assign or predict the target value of a new instance by checking all respective attributes and their values against those seen in the decision tree model. We carried out online pruning (performed during tree creation) to handle outliers and identify overfitting and subsets of instances that were poorly defined. The results are shown in Tables 3 and 4. The proposed decision tree approach exhibited accuracy levels of 70.87% and 64.28% in predicting delay occurrence in the US and Iranian networks, respectively. The method achieved accuracy levels of 63.89% and 70.79% in predicting delay magnitude in the US and Iranian networks, respectively. Class prediction shows a fraction of retrieved instances that are relevant. The precision of class prediction via decision tree ranged from 48.69% to 68.43% and 66.44% to 72.66% for various magnitudes of delays in the US and Iranian networks, respectively. Recall is the fraction of relevant instances that are retrieved. Class recall via decision tree ranged from 53.86% to 73.19% and 55.98% to 91.24% for the US and Iranian networks, respectively.

Table 4. Accuracy of flight delay occurrence and magnitude prediction.

Network	The US network		Iran network	
Prediction method	Accuracy of delay occurrence prediction (%)	Accuracy of delay magnitude prediction (%)	Accuracy of delay occurrence prediction (%)	Accuracy of delay magnitude prediction (%)
J48 pruned tree approach	64.28	63.89	70.87	70.79
Cluster classification approach	62.27	61.37	69.63	68.84
Hybrid approach	71.39	70.16	76.44	75.93
Bayesian approach	61.35	60.78	70.17	69.81
Random forest approach	67.43	66.27	72.11	71.23

Table 5. Accuracy of flight delay magnitude prediction in different classes.

Network		The US network		Iran network	
Prediction method	Class ranges (minute)	Accuracy of delay magnitude prediction (%)		Accuracy of delay magnitude prediction (%)	
		Class precision	Class recall	Class precision	Class recall
J48 pruned tree approach	0-15	68.43	73.19	71.57	60.11
	15-60	59.21	60.10	66.44	55.98
	> 60	48.69	53.86	72.66	91.24
Cluster classification approach	0-15	58.15	64.19	68.75	55.31
	15-60	53.27	58.49	62.63	56.36
	> 60	59.51	50.07	72.26	88.99
Hybrid approach	0-15	79.37	82.18	75.96	68.70
	15-60	60.28	68.81	74.00	64.35
	> 60	62.23	60.19	77.01	90.54
Bayesian approach	0-15	56.54	62.57	76.35	51.69
	15-60	51.94	56.95	64.82	58.02
	> 60	58.36	48.61	69.74	94.01
Random forest approach	0-15	69.18	73.21	67.26	66.09
	15-60	58.76	60.48	65.00	56.52
	> 60	51.37	52.34	75.15	83.78

5.2. Cluster classification

The results of cluster classification via the K-means algorithm are displayed in Tables 4 and 5. The algorithm performed with accuracies of 62.27% and 69.63% in predicting delay occurrence in the US and Iranian networks, respectively. It performed with accuracies of 61.37% and 68.84% in predicting delay magnitude in the US and Iranian networks, respectively. Class precision via the algorithm ranged from 53.27% to 59.51% and 62.63% to 72.26% for various magnitudes of delay in the US and Iranian networks, respectively. Class recall ranged from 55.31% to 88.99% and 50.07% to 64.19% for the US and Iranian networks, respectively. As can be seen, the decision tree generated results that were superior to those of the cluster classification.

5.3. Hybrid approach

In this approach, we combined the decision tree with cluster classification approach (i.e., the J48 decision tree algorithm and the K-means algorithm). The

results are provided in Tables 4 and 5. The hybrid approach exhibited accuracy levels of 71.39% and 76.44% in delay occurrence prediction for the US and Iranian networks, respectively. Its accuracy levels were 70.16% and 75.93% in predicting delay magnitude in the US and Iranian networks, respectively. Class precision was in the ranges of 60.19% to 82.18% and 64.35% to 90.54% for various magnitudes of delay in the US and Iranian networks, respectively. Class recall fell in the ranges of 60.28% to 79.37% and 74.00% to 77.01% for the US and Iranian networks, respectively. The results of the hybrid approach are superior to those of all the other methods.

5.4. Bayesian classification

The Bayesian classification results are also shown Tables 4 and 5. Its accuracy levels were 61.35% and 70.17% in predicting delay occurrence and 60.78% and 69.81% in predicting delay magnitude in the US and Iranian networks, respectively. Class precision ranged from 51.94% to 58.36% and 64.82% to 76.35% for various magnitudes of delays, and class recall ranged from 48.61% to 62.57% and 51.69% to 94.01% for the US and Iranian networks, respectively.

5.5. Random forest method

The results of the random forest algorithm are listed Tables 4 and 5. The algorithm achieved accuracy levels of 67.43% and 72.11% in predicting delay occurrence and 66.27% and 71.23% in predicting delay magnitude in the US and Iranian networks, respectively. Class precision was in the ranges of 51.37% to 69.18% and 65.00% to 75.15% for various magnitudes of delays, and class recall fell in the ranges of 52.34% to 73.21% and 56.52% to 83.78% for the US and Iranian networks, respectively.

5.6. Discussion

We compared the results of the proposed approaches and verified the findings on the basis of parameters such as class precision and class recall. The outputs of all the methods are shown in Tables 6 and 7 and Figures 11 and 12 for the US and Iranian networks, respectively. Among the approaches, the hybrid method produced the best results, having minimum class recall of 60.19% for the US network and 64.35% for the Iranian network. Class recall is an important parameter that determines the reliability of results. The accuracy levels of the hybrid approach were 71.39% and 76.44% in predicting delay occurrence and 70.16% and 75.93% in predicting

Table 6. Comparison of various approaches (the US network).

Approach	Delay occurrence	The magnitude of delay				
	Accuracy (%)	Accuracy (%)	Class recall (%)		Class precision (%)	
			Min	Max	Min	Max
Decision tree	64.28	63.89	53.86	73.19	48.69	68.43
Cluster classification	62.27	61.37	50.07	64.19	53.27	59.51
Hybrid method (decision tree combined with cluster classification)	**71.39**	**70.16**	**60.19**	**82.18**	**60.28**	**79.37**
Bayesian approach	61.35	60.78	48.61	62.57	51.94	58.36
Random forest	67.43	66.27	52.34	73.21	51.37	69.18

Table 7. Comparison of various approaches (Iran network).

Approach	Delay occurrence	The magnitude of delay				
	Accuracy (%)	Accuracy (%)	Class recall (%)		Class precision (%)	
			Min	Max	Min	Max
Decision tree	70.87	70.79	55.98	91.24	66.44	72.66
Cluster classification	69.63	68.84	55.31	88.99	62.63	72.26
Hybrid method (decision tree combined with cluster classification)	**76.44**	**75.93**	**64.35**	**94.54**	**74.00**	**77.01**
Bayesian approach	70.17	69.81	51.69	94.01	64.82	76.35
Random forest	72.11	71.23	56.52	83.78	65.00	75.15

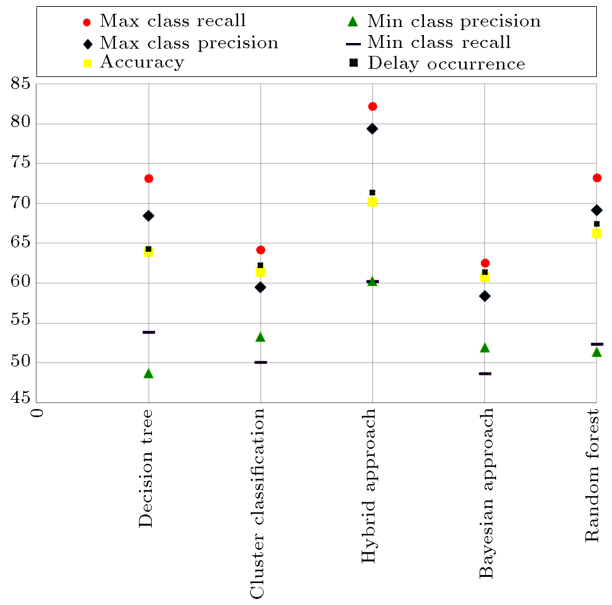

Figure 11. Comparison of various approaches (the US network).

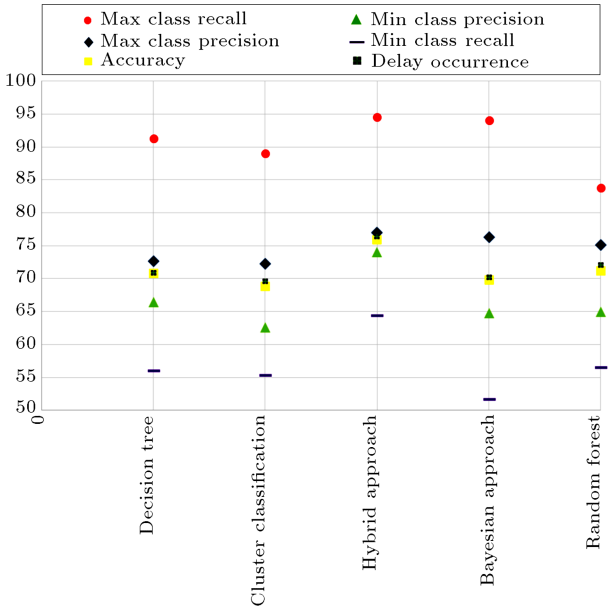

Figure 12. Comparison of various approaches (Iran network).

delay magnitude in the US and Iranian networks, respectively. Class precision ranged from 60.28% to 79.37% and 74.00% to 77.01% for the US and Iranian networks, respectively. These values were the highest among the approaches. Although we also modeled a public database of the US network, the Iranian network is special because of the situation of the country. Our results showed that the proposed model could predict delays occurring the US network at an accuracy of 70%, which was a reasonable level. It could predict delays in the Iranian network at an accuracy of 75%.

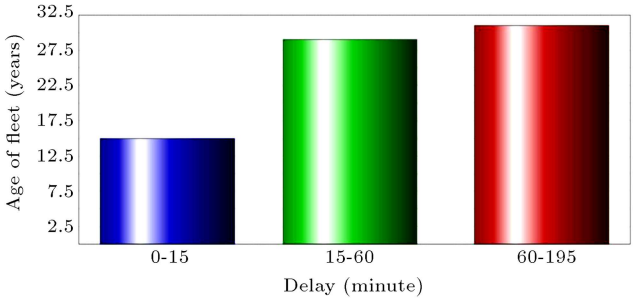

Figure 13. Predicting effect of the age of the fleet on delays.

Moreover, weather conditions are very important in delays in the US network, whereas fleet age and fleet type are more significant in the Iranian network. As previously indicated, fleet age and fleet type in Iran are of lower quality than the standard because of sanctions in previous years and maintenance issues. The effects of fleet age on the Iranian network are illustrated in Figure 13. The FDP results indicated that older airplanes experienced more delays (Figure 8).

6. Conclusions and recommendations

FDP methods, namely decision tree, cluster, Bayesian, random forest, and hybrid classification, were proposed in this research. These approaches were examined on the basis of real datasets on US and Iranian flight networks. The results indicated that the hybrid approach exhibited a performance superior to those of the other methods and was therefore adopted as the FDP model. Parameters such as fleet age and aircraft type exert strong effects on flight delays in the Iranian network, whereas weather conditions strongly influence flight delays in the US network. The accuracy levels of the hybrid approach were 71.39% and 76.44% in predicting delay occurrence and 70.16% and 75.93% in predicting delay magnitude in the US and Iranian networks, respectively. These results may be of interest to airlines that want to implement measures for preventing delay propagation, especially those based in developing countries, such as Iran.

For the future studies, researchers can implement other exciting data mining methods and compare the results. The proposed combined method of delay anticipation and its results can also be further explored in other studies. For example, combing the hybrid method with robust flight scheduling shows potential as an interesting research direction.

References

1. Barnhart, C. and Smith, B. "Quantitative problem solving methods in the airline industry", *International Series in Operation Research & Management Science*, Springer Science + Business Media, LLC (2012).

2. Ball, M., Barnhart, C., and Drenser, M., *Total Delay Impact Study*, The National Center of Excellence for Aviation Operation Research (2010).
3. Bazargan, M., *Airline Operations and Scheduling*, MPG Book Group, 2th Ed., UK (2010).
4. Barnhart, C. and Amy, C. "Airline schedule planning: accomplishments and opportunities", *Manufacturing & Service Operations Management*, **6**(1), pp. 3-22 (2004).
5. Gopalakrishnan, B. and Johnson, E.L. "Airline crew scheduling: state-of-the-art", *Annals of Operations Research*, **140**, pp. 305-337 (2005).
6. Sherali, H.D., Bish. E.K., and Zhu, X. "Airline fleet assignment concepts, models, and algorithms", *European Journal of Operational Research*, **172**, pp. 1-30 (2005).
7. Listes, O. and Dekker, R. "A scenario aggregation-based approach for determining a robust airline fleet composition for dynamic capacity allocation", *Transportation Science*, **39**(3), pp. 367-382 (2005).
8. Akartunal, K., Boland, N., Evans, I., Wallace, M., and Waterer, H. "Airline planning benchmark problems-part II: passenger groups, utility and demand allocation", *Computers & Operations Research*, **40**, pp. 793-804 (2013).
9. Kohla, N., Larsen, A., Larsen, J., Ross, A., and Tiourine S. "Airline disruption management-perspectives, experiences and outlook", *Journal of Air Transport Management*, **13**, pp. 149-162 (2007).
10. Clausena, J., Larsen, A., Larsen, J., and Rezanova, N.J. "Disruption management in the airline industry-concepts, models and methods", *Computers & Operations Research*, **37**, pp. 809- 821 (2010).
11. http://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp.
12. Transportation Research Board "Defining and measuring aircraft delay and airport capacity thresholds", ACRP Report 104 (2014).
13. Tu, Y., Ball, M., and Jank, W. "Estimating flight departure delay distributions -a statistical approach with long-term trend and short-term pattern", *Journal of the American Statistical Association*, **103**, pp. 112-125 (2008).
14. Mueller, E.R. and Chatterji, G.B. "Analysis of aircraft arrival and departure delay characteristics", *Proceeding of the 2th AIAA's Aircraft Technology, Integration, and Operations (ATIO) Conference*, Los Angeles, California, USA (2002).
15. Avijit, M., Lovell, D.J., Ball, M.O., Odoni, A.R., and Zerbib, G. "Modeling delays and cancellation probabilities to support strategic simulations", *Proceedings of the 6th Air Traffic Management Research and Development Seminar*, Baltimore, MD, USA (2005).
16. Sridhar, B., Wang, Y., Klein, A., and Jehlen, R. "Modeling flight delays and cancellations at the national, regional and airport levels in the United States", *Proceedings of the 9th Air Traffic Management Research and Development Seminar*, Berlin, Germany (2011).
17. Lu, Z., *Alarming Large Scale of Flight Delays: An Application of Machine Learning*, Machine Learning. In Tech publishing, pp. 239-250 (2010).
18. Lu, Z., Wang, J., and Zheng, G. "A new method to alarm large scale of flights delay based on machine learning, in knowledge acquisition and modeling", *KAM '08. International Symposium on*, pp. 589-592 (Dec. 2008).
19. Bolaños, M.E. and Murphy, D. "How much delay does New York inject into the national airspace system? A graph theory analysis", *Proceeding of the 11th AIAA's Aircraft Technology, Integration, and Operations (ATIO) Conference*; Los Angeles, California, USA (2013).
20. Rebollo, J.J. and Balakrishnan, H. "Characterization and prediction of air traffic delays", *Transportation Research Part C*, **44**, pp. 231-241 (2014).
21. Oza, S., Sharma, S., Sangoi, H., Raut, R., and Kotak, V.C. "Flight delay prediction system using weighted multiple linear regression", *International Journal of Engineering and Computer Science*, **4**(4), pp. 11668-11677 (2015).
22. Allan, S.S., Gaddy, S.G., and Evans, J.E., *Delay Causality and Reduction at the New York City Airport Using Terminal Weather Information*, Massachusett Institute of Technology (2011).
23. Wu, C. "Inherent delays and operational of airline schedules", *Journal of Air Transportation Management*, **11**(4), pp. 273-282 (2005).
24. Wang, P., Schaefer, L., and Wojcik, L. "Flight connections and their impacts on delay propagation", Technical Report, MITRE (2003).
25. Janic, M. "Modeling the large scale disruption of an airline network", *Journal of Transportation Engineering*, **131**(4), pp. 249-260 (2005).
26. Hsiao, C. and Hansen, M. "An econometric analysis of us airline flight delays with time-of-day effects", *Proceedings of TRB 2006 Annual Meeting* (2006).
27. Rupp, N. "Investigating the causes of flight delays", Working paper, Department of Economics, East Carolina University (2007).
28. Boswell, S.B. and Evans, J.E. "Analysis of downstream impacts of air traffic delay", Lincoln Laboratory, Massachusetts Institute of Technology (1997).
29. Chen, H., Wang, J., and Yan, X. "A fuzzy support vector machine with weighted margin for flight delay early warning", In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD '08. Fifth International Conference on*, **3**, PP. 331-335 (Oct. 2008).
30. Liao, S.H., Chu, P.H., and Hsiao, P.Y. "Data mining techniques and applications - A decade review from 2000 to 2011", *Expert Systems with Applications*, **39**, pp. 11303-11311 (2012).

31. Vehtari, A. and Ojanen, J. "A survey of Bayesian predictive methods for model assessment, selection and comparison", *Statistics Surveys*, **6**, pp. 142-228 (2012).
32. Han, J., Kamber, M., and Pei, J., *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers, 3th Ed. (2012).
33. Breiman, L. "Random forests", *Machine Learning*, **45**(1), pp. 5-35 (2001).
34. Singh, K., Malik, D., and Sharma, N. "Evolving limitations in K-means algorithm in data mining and their removal", *IJCEM International Journal of Computational Engineering & Management*, **12**(1), pp. 105-109 (2011).
35. Abbas, O.A. "Comparisons between data clustering algorithms", *The International Arab Journal of Information Technology*, **5**(3), pp. 320-325 (2008).
36. https://www.wunderground.com.
37. Liu, Y., Yu, X., and Huang, J.X. "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets", *Information Processing & Management*, **47**(4), pp. 617-631 (2011).
38. Mirkin, B. "Core concepts in data analysis: summarization, correlation and visualization", Springer: Verlag London Limited: Available from Researchgate: http://www.researchgate.net/ profile/Boris_Mirkin/ publication/232282057_Core_Concepts_in_Data_ Analysis_Summarization_Correlation_and_Visualization/ links/0912f51090564b6e36000000.pdf (2011).
39. Maimon, O. and Rokachm, L. *Data Mining and Knowledge Discovery Handbook*, Springer, pp. 853-867, ISBN: 978-0-387-24435-8 (Print), 978-0-387-25465-4 (Online) (2005).
40. https://rapidminer.com/
41. Sharma, T.C., Jain, M., and abad, F. "WEKA approach for comparative study of classification algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, **2**(4), pp. 1925-1931 (2013).

Biographies

Hassan Khaksar is PhD candidate in the Department of Transportation Engineering and Planning, School of Civil Engineering, Iran University of Science and Technology (IUST). He received BSc in 2005 and MSc in 2007 in Civil Engineering from IUST. He is doing the PhD thesis entitled "The Optimum Airline Planning Model Based on Disruption Management and Robust Planning." This research focuses on predicting and reduction of delays of airlines. He has published more than 30 scientific paper in ISI journals and international conferences. His research interests are airline planning, airport managing, data mining, and operation research.

Abdolrreza Sheikholeslami is Associate Professor in the Department of Transportation Engineering and Planning, School of Civil Engineering, Iran University of Science and Technology (IUST). He received his BSc and MSc degrees from IUST and his PhD degree in Transportation Engineering and Planning from the same university in 2006. His research interests are transportation planning, transportation economics, airline scheduling, disruption management, data mining, railway scheduling, mathematical models, and optimization techniques for designing transportation systems. He was elected as the top student in the technical and engineering group in Iran in 2001. He is experienced with more than 20 years of teaching at IUST. He has published more than 80 scientific papers in ISI journals and international conferences.