Modeling Air Travel Choice Behavior with Mixed Kernel Density Estimations

Zhenni Feng
Shanghai Jiao Tong University
Shanghai, China
zhennifeng@sjtu.edu.cn

Yanmin Zhu*
Shanghai Jiao Tong University
Shanghai, China
yzhu@sjtu.edu.cn

Jian Cao
Shanghai Jiao Tong University
Shanghai, China
cao-jian@sjtu.edu.cn

ABSTRACT

Understanding air travel choice behavior of air passengers is of great significance for various purposes such as travel demand prediction and trip recommendation. Existing approaches based on surveys can only provide aggregate level air travel choice behavior of passengers and they fail to provide comprehensive information for personalized services. In this paper we focus on modeling individual level air travel choice behavior of passengers, which is valuable for recommendations and personalized services. We employ a probabilistic model to represent individual level air travel choice behavior based on a large dataset of historical booking records, leveraging several key factors, such as takeoff time, arrival time, elapsed time between reservation and takeoff, price, and seat class. However, each passenger has only a limited number of historical booking records, causing a serious data sparsity problem. To this end, we propose a mixed kernel density estimation (mix-KDE) approach for each passenger with a mixture model that combines probabilistic estimation of both regularity of the individual himself and social conformity of similar passengers. The proposed model is trained and evaluated via the expectation-maximization (EM) algorithm with a huge dataset of booking records of over 10 million air passengers from a popular online travel agency in China. Experimental results demonstrate that our mix-KDE approach outperforms the Gaussian mixture model (GMM) and the simple kernel density estimation in the presence of the sparsity issue.

CCS Concepts

•Mathematics of computing → Kernel density estimators; •Computing methodologies → Mixture models; •Information systems → Recommender systems;

*Corresponding author.

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018671

Keywords

Air travel choice behavior; individual level modeling; mixed kernel density estimation

1. INTRODUCTION

Nowadays, it is convenient to travel by plane for taking holidays, visiting friends or business purposes. It is reported by Civic Aviation Administer of China that there are over 38 million people travel by air during May 2016 in China [1], which is almost 10% increase compared with the same month of 2015. Understanding air travel choice behavior of air passengers is of great significance for various applications such as travel demand prediction, trip planning and recommendation.

Existing approaches for understanding air travel choice behavior typically rely on revealed preference and stated preference data from passenger surveys [2, 6]. These surveys help to investigate air passengers' preference and to estimate a few valuable metrics, e.g., value of travel time savings, willingness-to-pay. Such approaches, however, can only provide aggregate level air travel choice behavior of passengers, which fails to provide comprehensive information about individuals.

With the widespread prevalence of the Internet and online payment tools, people usually reserve their flights in advance at online travel agencies, such as www.expedia.com, www.ctrip.com, and www.qunar.com. When a passenger plans to reserve a flight online, he often considers several key attributes of a flight (e.g., airline, aircraft type, takeoff time) according to personal preference or background (e.g., employment). The online flight reservation system then stores the booking record, which contains fare, destination, airline carrier choice, seat, aircraft type, flight time, number of connections, takeoff time and arrival time of the trip. With an ever-increasing dataset of flight booking records available in these online travel agencies, there is an unprecedented opportunity to understand individual level air travel choice behavior.

In this paper, we focus on *understanding individual level air travel choice behavior of air passengers which greatly impacts the way how one books flights* upon a huge dataset of online reservation records of flights collected by a popular online travel agency in China. More specifically, this paper aims at developing an accurate probabilistic model for each air passenger, characterizing several key factors, such as takeoff and arrival time, elapsed time between reservation and takeoff, price, and seat class.

However, modeling air travel choice behavior at the indi-

vidual level is faced with several challenges. *Firstly*, it is a multi-variate modeling problem since the choice behavior is characterized by a number of key factors, e.g., fare, flight time, etc. *Secondly*, little study exists for modeling the air travel choice behavior at the individual level. Widely used parametric distribution functions, e.g., Gaussian distribution, Poisson distribution, are inappropriate for this choice behavior modeling. *Thirdly*, each passenger has only a limited number of historical booking records. It is observed with our dataset that a passenger has on average as few as 5 booking records per year, which leads to a serious data sparsity problem.

To this end, *we address the problem based on a mixed kernel density estimation approach that leverages not only his own regularity but also social conformity with "similar" air passengers.* Regularity of an individual passenger is regarded as his own preference on air travel choice behavior while social conformity indicates that one's choice behavior is influenced by other "similar" passengers possibly due to their similar backgrounds or experiences. The proposed method is inspired by the intuition above: 1) For each air passenger, we first employ a non-parametric density estimation method to model his air travel choice behavior, in order to grasp his own regularity hidden in his own historical air travel trips. 2) To address the data sparsity problem, we further introduce a probabilistic estimation upon "similar" passengers to represent the conformity property, also based on kernel density estimation. This part is called coarse-level estimation. 3) We propose mix-KDE that builds a probabilistic model which combines the individual's kernel density estimation and coarse-level estimation. The proposed method is then trained via the expectation-maximization (EM) algorithm and evaluated with a huge dataset of booking records of over 10 million users of a popular online travel agency in China. Main contributions of this paper are summarized as follows:

- We introduce the air travel choice modeling problem as a multi-variate modeling problem and then point out the data sparsity issue.
- We propose a novel mix-KDE model that utilizes both regularity and social conformity of air travel choice behavior and then train the model via a huge dataset of online reservation of flights.
- Experimental results demonstrate that our mix-KDE approach outperforms the GMM and the simple kernel density estimation in the presence of data sparsity. In addition, we further offer some visualization on the proposed model.

The rest of the paper is organized as follows. We describe the air travel choice behavior modeling problem in Section 2. In Section 3, we extract some related features in the dataset. Then, we discuss the proposed mix-KDE approach in Section 4 and 5. Then, experimental results are given in Section 6. Literature review is provided in Section 7. Finally, we conclude our paper in Section 8.

2. AIR TRAVEL CHOICE BEHAVIOR MODELING PROBLEM

In the section, we first explain the heterogeneity of passengers via several simple observations and then offer the problem description in the end of this section. Main notations and their meanings used in the paper are summarized in Table 1.

Table 1: Summary of notations.

Notation	Meaning
i	Passenger id, often used as subscript
k	Record number, e.g., the k-th record of a passenger, often used as subscript
$t^r_{i,k}, t^f_{i,k}, t^a_{i,k}$	Reservation time, takeoff time (flight time), arrival time
$c^d_{i,k}, c^a_{i,k}$	Departure city, arrival city
$p_{i,k}, \rho_{i,k}$	Price, price discount
$s_{i,k}$	Seat class, i.e., $\{Y$=economy, F=first class, C=business$\}$
a_i, g_i	Age, gender of a passenger
f_i	Frequency of taking flights, i.e., average number of flights in a year
α_i	Elapsed time between reservation and takeoff, i.e., $t^f_{i,k} - t^r_{i,k}$
β_i	Time difference between two consecutive flights, i.e., $t^f_{i,k} - t^f_{i,k-1}$
$\mathbb{D}_i$	The set of reservation records of passenger i
$Day(\cdot), Hour(\cdot)$	Day of a date, hour of a day
$Days(\cdot)$	Number of days of time difference
$\overline{(\cdot)}$ or $avg(\cdot), std(\cdot)$	Average, standard deviation over a passenger
$Ent(\cdot)$	Entropy function, often used to measure diversity of a parameter
$Prop(\cdot)$	Proportion function, often used to measure proportion of a parameter

2.1 Heterogeneous air travel choice behaviors

Air passengers usually have their own preferences towards various reservation options (e.g., reservation time) or service attributes (e.g., airline, seat, aircraft type, price discount), all of which are recorded by their booking records. Diversity in air passengers' choice behavior is generally due to personal backgrounds (e.g., demographics, employment) and experiences (e.g., past trips).

We offer an example that explains different passengers have rather different preferences on various options with some statistical information. As shown in Table 2, each row offers a kind of statistical information of passengers (i.e., A, B and C), and average or standard deviation over randomly selected 10,000 passengers. In the first row, we can see that three passengers have various air travel frequencies, from 10 to 30 flights per year. In addition, passenger C shows smaller $\overline{\rho}$ and $\overline{\alpha}$, which indicates that passengers C prefers lower price discounts and reserves his tickets much later. In spite of more frequent flights of passenger C, time difference between two consecutive flights, $\overline{\beta_i}$, of passenger C is longer than that of passenger B as shown in the bottom row. This happens when passenger B takes most of his flights in a short time period while passenger C travels at more evenly distributed time points in a year.

Table 2: Comparison of statistical information.

	Passenger A	Passenger B	Passenger C	Avg_{10000}	Std_{10000}
f_i	10	20	30	14.219	5.340
$\overline{\rho_i}$	0.844	0.821	0.762	0.667	0.130
$\overline{\alpha_i}$	3.425	4.527	1.852	8.123	3.638
$\overline{\beta_i}$	15.786	8.697	11.691	21.643	6.676

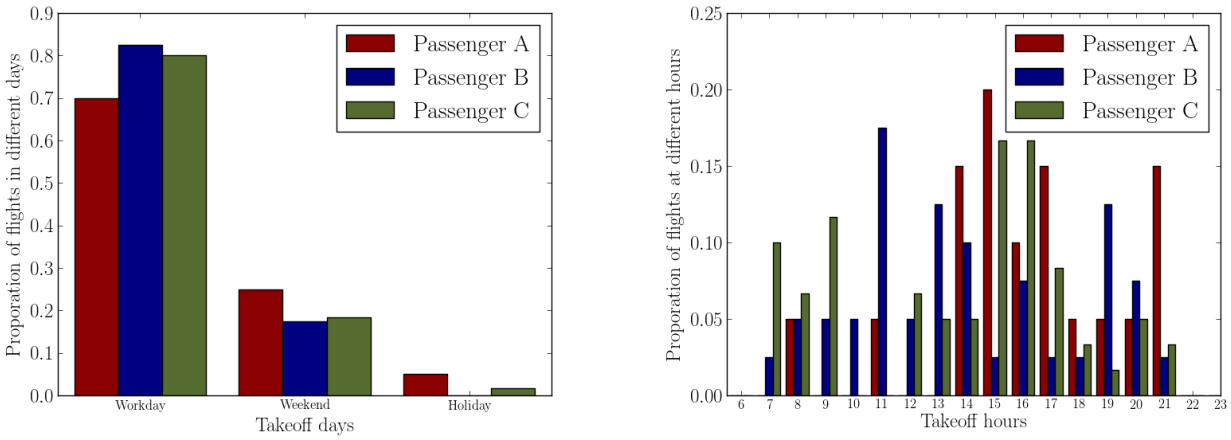

Figure 1: Comparison of preferences on takeoff days and takeoff hours.

Besides, their preferences on takeoff days and takeoff hours are displayed in Fig. 1. In the left figure, a takeoff day is classified into three types (the rationale of classification is shown in Subsection 3.2 later) and then the proportion of each type is computed for these passengers. The figure in the right reflects their preferences on different takeoff hours. Passenger A prefers taking a flight in the afternoon while passenger B shows much more preferences on flights at 11 a.m. Different distributions on takeoff days or takeoff hours are probably due to their different employments and professions.

From the observations above, we can conclude that different air passengers show different preferences upon each option and thus have different air travel choice behaviors. Consequently, we aim to model the air travel choice behavior at the individual level, i.e., proposing a probabilistic model for each air passenger.

2.2 Problem description

Definition (**Personalized air travel choice behavior modeling problem**) Suppose the dataset of booking records $\mathbb{D} = \cup_i \mathbb{D}_i$, where i is the identification of one air passenger, collects N air passengers' historical air traveling records. Each record, $d_i \in \mathbb{D}_i$, contains a set of factors, $\mathbf{x} = (x^{(1)}, x^{(2)}, \cdots, x^{(m)})$. The air travel choice behavior modeling problem is to derive a probability density model, $\mathcal{M}_i(\mathbf{x})$, for each air passenger i that fits the data $\mathbb{D}_i$ as much as possible.

3. FEATURE EXTRACTION

In the section, we describe several key factors that have great impact on air travel choice behavior of an air passenger. We summarize these factors into three aspects, i.e., reservation factors, flights factors and passenger factors.

3.1 Reservation factors

When a passenger books a ticket on a travel agency for his air trip, some reservation factors that reflect his preference are introduced, e.g., elapsed time between reservation and takeoff α, and time difference between two consecutive flights β. To explore these factors, we randomly choose 10,000 passengers and derive statistical results upon their historical flights.

We plot the empirical PDF of α in the left picture of Fig. 2. We can observe that people have various preferences on their reservation time, ranging from less than a day to more than a month. We further observe that most flights are reserved $0.85 \times 24 \approx 20$ hours in advance of takeoff time. The average number of days between reservation time and takeoff time varies from passenger to passenger. For example, price-sensitive people are likely to book their flights early in order to get cheap tickets while business people usually book flights late due to urgent business trips.

Another plot in Fig. 2 shows distribution of time differences between two consecutive flights. The most popular β is 2.7 days which is consistent with round-trip tickets during short holidays of three days such as Labor's Day. However, longer elapsed time (e.g., longer than a week) between two consecutive flights happens since they belongs to two different air trips.

3.2 Flight factors

Takeoff day is an important factor when booking a flight. We plot a heat map of the number of air passengers in different days from December 1, 2013 to November 30, 2014 in

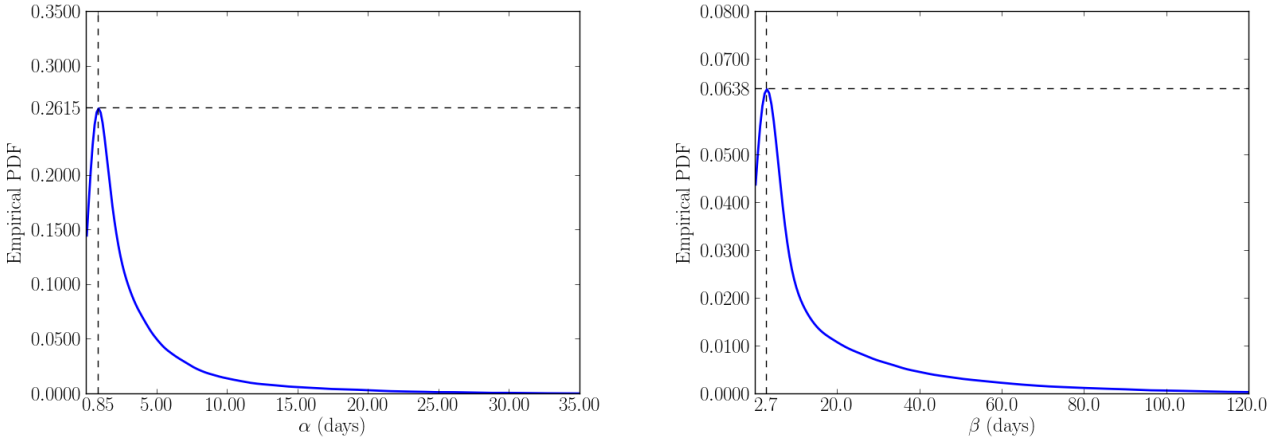

Figure 2: Empirical PDF of α and β.

Fig. 3. It is clear that there are four peaks during the whole year. The first two peaks appear at late January and early February. The period between the two peaks is corresponding to the Chinese Lunar New Year holidays. People often travel to their hometowns for family reunion before holidays and then go back to the cities where they work after holidays. Similarly, another important festival in China, the National Day holiday (usually from October 1 to October 7) generates another two peaks. During the National Day holiday, people usually go for sightseeing as well as visit their families. Apart from the most two important holidays, there were also some other holidays, e.g., Tomb-sweeping Day at early April, which also saw a higher density of passengers. In addition, we can also see from the figure that there is a periodical pattern with a period of 7 days. For instance, during December 2013, in the day of 7th, 14th, 21st and 28th show lighter color than that of others, indicating much smaller numbers of passengers in these days. Above all, the takeoff day factor is reasonably described by a categorical value of {H=Holiday, W=Workday and E=weekEnd (excluding holidays)}.

Price discount is another factor that affects air travel choice behavior of passengers. Empirical PDF of discounts based on historical records of randomly selected 10,000 passengers is plotted in Fig. 4. The most common discount is more than 0.9 and discounts are usually larger than 0.4. The PDF line has local peaks at "integral points" because price discounts are generally given by integral numbers, e.g., at a 30% discount instead of 21% off.

3.3 Passenger factors

Impacting factors related with passengers themselves may also affect their air travel choice behavior, e.g., age and gender. In Fig. 5, we plot proportion of flights with respect to passengers' ages. It is obvious that age is a significantly important factor since it is closely related with people' employment, financial conditions or leisure time.

To sum up, these features are extracted as impacting factors in air travel choice behavior models, including α, Day(takeoff time), $Hour$(takeoff time), seat class and price discount. Note that passenger factors are not included because we aim to construct individual behavior models. Passenger factors are employed to model social conformity in Section 5 later.

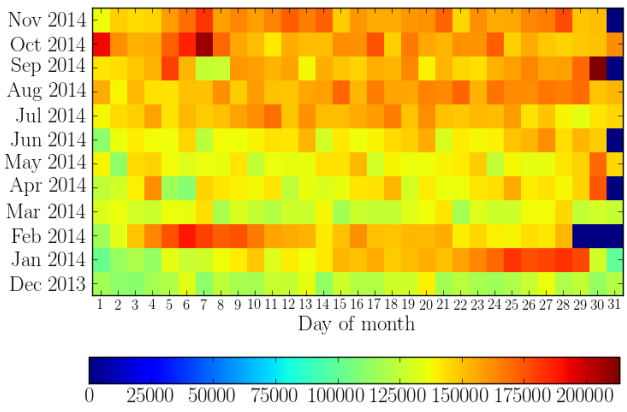

Figure 3: The numbers of passengers at different days in different months.

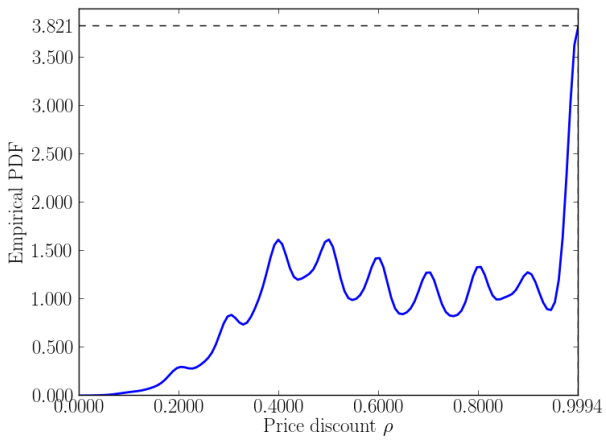

Figure 4: Empirical distribution of price discount.

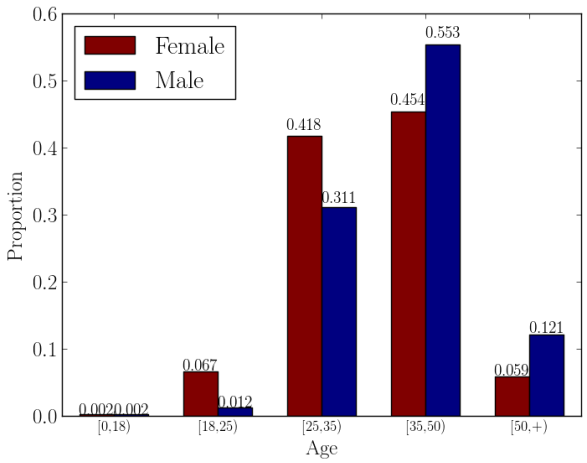

Figure 5: Empirical distribution of flights in different groups of ages.

4. INDIVIDUAL LEVEL AIR TRAVEL CHOICE BEHAVIOR MODEL

In the section, we introduce kernel density estimation (KDE) to model the air travel choice behavior for each passenger. Firstly, the preliminaries of KDE are given. Then, we perform a data-driven bandwidth determination scheme to derive a density model for each passenger.

4.1 Kernel density estimation

Kernel density estimation is a kind of non-parametric estimation that attempts to estimate the underlying density directly from data without assuming a particular form of the underlying distribution. The histogram is regarded as the simplest kind of non-parametric density estimation. However, the histogram still has two "parameters" to determine: bin width and starting point of the first bin. What is worse, the histogram ignores the importance of each individual sample and only counts the number of samples that falls in a particular bin. To overcome the drawbacks of histogram, a smoothing kernel function is introduced to replacing the fixed-width bins, thus introducing kernel density estimation. The kernel function is a probability distribution function $G(u)$ that satisfies the following condition, i.e.,

$$\int_{x \in R^D} G(x)dx = 1, \tag{1}$$

where D is the dimension of x.

Thus, the kernel density estimation is as follows:

$$g(x) \propto \sum_k G(\frac{x - x^{(k)}}{h}), \tag{2}$$

where h is called bandwidth.

4.2 Bandwidth selection

The selection of bandwidth is a crucial problem in density estimation. Either too small or too large bandwidth would degrade the performance of estimation result.

In our problem, we employ the kernel density estimation as follows. Generally, we choose the commonly used Gaussian function as the kernel function. For each air passenger i, the model $g_i(x)$ is generated by its own booking records $\mathbb{D}_i$. Instead of choosing a numerical bandwidth parameter using a commonly known rule [13], we employ a data-driven bandwidth selection method, i.e., finding the best bandwidth that maximizing likelihood via cross validation. In the cross validation process, the model is fitted to part of the data, and then it is evaluated by a specific metric that measures how well this model fits the remaining data. The bandwidth that fits data best is chosen. For each air passenger i, suppose the selected bandwidth is $\hat{h_i}$, its density estimation is

$$g_i(x) \propto \sum_{x^{(k)} \in \mathbb{D}_i} G(\frac{x - x^{(k)}}{\hat{h_i}}). \tag{3}$$

The coefficients in $g_i(x)$ is easy to compute by normalizing the summation of $g_i(x)$ to 1.

5. TACKLING SPARSITY PROBLEM WITH MIXTURE MODELS

In the section, we point out the data sparsity problem and address the problem via a mix-KDE approach that builds a mixture model for each passenger. Each component in the mixture model is either the individual's kernel density estimation or coarse-level estimation from a population of individuals sharing similar behavior with the passenger.

5.1 Sparsity problem

Since a lot of air passengers only have a few records (less than 30 records during 2 years), we regard these passengers as inactive ones. For example, if a passenger only bought 3 tickets in two years in the dataset (which is probably because the dataset only covers a small part of her/his air trips), the air passenger is regarded as inactive. Since the air travel choice behavior of inactive passengers is rather random and noisy, we only consider active ones in the paper. We plot the number of records for active air passengers in Fig. 6. We can see that a large proportion of passengers have less than 50 records. Therefore, even for active passengers, these records are not sufficient to generate accurate density models. This is because only thirty or fifty points are not enough to generate an accurate density model. We call this problem the sparsity problem in the paper.

5.2 Measuring difference between passengers

We extract several statistics from historical booking records, in order to explain the differences between each pair of individual air passengers. We call these statistics *profile vector* of an air passenger and all fields included in the profile vector are listed in Table 3.

Commonly known distance measures such as Euclidean distance can be employed to evaluate the differences between profile vectors. Furthermore, profile vectors are normalized at each dimension for simplicity. For example, profile vectors of three air passengers are given in Table 4. From Table 4, we can see that passenger A and passenger B prefer booking economy seats than passenger C, and B usually plans his trips earlier than the other two. The pairwise distance is then computed according to Euclidean distance based on normalized profile vectors which is shown in Table. 5. The distance measures the difference between air passengers. The longer the distance is, the more difference the pair

Table 3: Fields in the profile vector

Name	Notation	Meaning
Travel frequency	f_i	Number of air trips per year
Seat preference	$Prop(s_{i,\cdot} = Y)$	Proportion of taking economy seats
Travel time preference	$Prop(t_{i,\cdot} = W)$	Proportion of traveling at workdays
Travel plan preference	$\overline{\alpha_{i,\cdot}}$	Average days between reservation time and takeoff time
Travel duration preference	$\overline{\beta_{i,\cdot}}$	Average days between current flight and last flight
Diversity of destinations	$Ent(c^a_{i,\cdot})$	Information entropy for destinations
Age	a_i	Age of a passenger

has. The results show that B and C are the most similar among all pairs.

We further offer a scatter plot of profile vectors of 10,000 air passengers in Fig. 7. The seat preference and travel time preference are chosen to plot scatter points. The density of each dimension is also shown in subplots of Fig. 7. From the figure we can see that a large proportion of people prefer economy seats. However, the preference on travel time is more diverse than seat preference.

5.3 Mixture models

Finite mixture models are a family of probability distribution functions of the form

$$\mathcal{M}(\mathbf{x}; \mathbf{p}, \theta) = \sum_{k=1}^{c} p_k g_k(\mathbf{x}; \theta_k), \tag{4}$$

where $\mathbf{x}$ is a d-dimensional random variable, $\mathbf{p} = [p_1, p_2, \cdots, p_c]$, and $\theta = [\theta_1, \theta_2, \cdots, \theta_c]$, with p_k being the mixture proportions, g_k being the component densities, with g_k parameterized by θ_k, c is the number of components. The mixture proportions are non-negative and $\sum_{k=1}^{c} p_k = 1$.

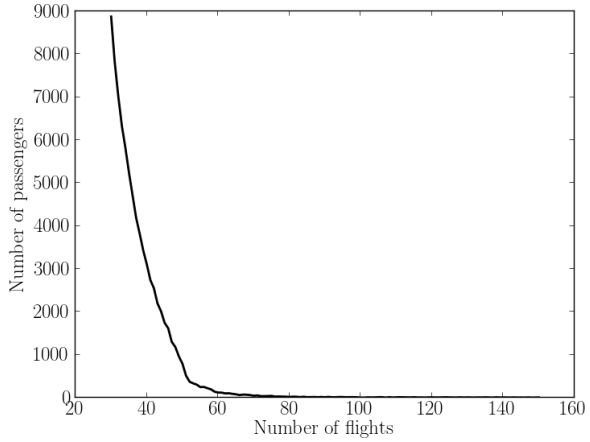

Figure 6: Number of air passengers vs. Number of flights (It only shows active passengers).

Table 4: Examples of profile vectors

ID	Profile vector	Normalized profile vector
A	(10.0, 1.0, 0.7, 3.35, 15.628, 1.4, 30)	(0.167, 1.0, 0.7, 0.056, 0.781, 0.267, 0.429)
B	(20.0, 1.0, 0.825, 4.525, 8.643, 1.475, 48)	(0.333, 1.0, 0.825, 0.075, 0.432, 0.281, 0.686)
C	(30.0, 0.933, 0.8, 1.9, 11.647, 0.766, 45)	(0.5, 0.933, 0.8, 0.032, 0.582, 0.146, 0.643)

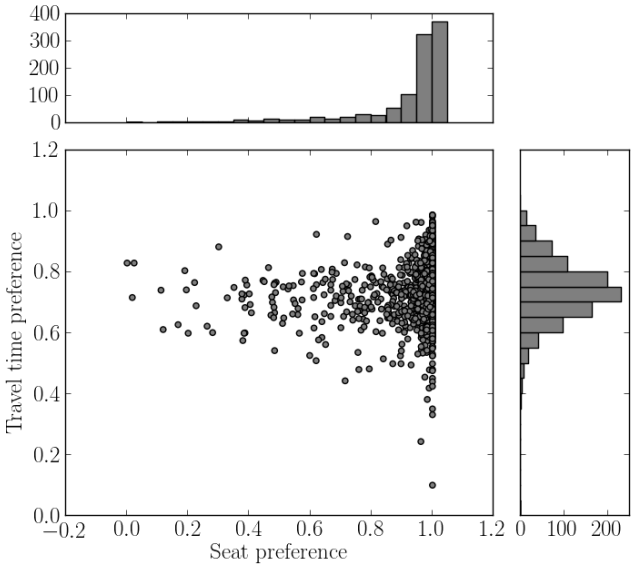

Figure 7: A scatter plot of profile vectors with only two fields of seat preference and travel time preference. The top and the right subplots show the density of each dimension, respectively.

Apart from regularity of each passenger modeled in Section 4, we introduce the idea of social conformity to address the sparsity problem. Specifically, air travel choice behavior is correlated with other "similar" passengers. Similar passengers can simply be interpreted as passengers with similar preferences, e.g., living in the same city, preferring a specific airline, etc. The aforementioned profile vectors can be employed to choose similar passengers. Then, the influence of similar passengers constitutes other components in the mixture models.

In our method, we determine the components via the distance between passengers. Consider a 2-component mixture model of passenger i, where the first component is generated

Table 5: Examples of distance of profile vectors

Pair of passengers	Distance
(A,B)	0.481
(A,C)	0.475
(B,C)	0.278

by the booking records of his own, the other component is generated by the records from other similar passengers. For more complicated models with more components, e.g., 3-component mix-KDE, all the other component are generated by a subset of similar air passengers. More specifically, the second component is generated with passengers those have a distance less than a threshold τ_1, and the third component is generated by the corresponding records of passengers with a distance not less than τ_1 and less than τ_2. Suppose the components are generated by $\mathbb{D}^1(=\mathbb{D}_i), \mathbb{D}^2, \cdots, \mathbb{D}^c$. The mixture models are written as follows:

$$\mathcal{M}_i(\mathbf{x}) = \sum_{k=1}^{c} p_k g_k(\mathbf{x}; \theta_k | \mathbb{D}^k), \tag{5}$$

where $g_k(\mathbf{x}; \theta_k | \mathbb{D}^k)$ denotes g_k is generated by dataset $\mathbb{D}^k$, c is the number of components.

5.4 Model training

Each component is solved by the kernel density estimation, i.e., g_k parameterized by θ_k is generated by Equation (3). Therefore, undetermined parameters are the mixture proportions, $p_k, 1 \leq k \leq c$. Expectation-Maximization (EM) algorithm [17] is employed to solve the problem because it is simple to implement and converges quickly.

Suppose the observations used for training the models is $y_j, 1 \leq j \leq n$, where n is the number of observations used for training models. In the EM framework for our problem, we introduce unobservable component labels $z_{kj} (1 \leq k \leq c, 1 \leq j \leq n)$ as the "missing" data (hidden variables), where z_{kj} is defined as one or zero according to whether y_j belongs to kth component of the mixture models. The EM algorithm is an iterative procedure with expectation (E-step) and maximization (M-step). The details of the EM algorithm is shown in Algorithm 1.

Algorithm 1 The EM algorithm for training models

Input: $\mathbf{y} = \{y_j | 1 \leq j \leq n\}$, $\mathbf{g} = \{g_k | 1 \leq k \leq c\}$, ϵ
Output: $\mathbf{p} = \{p_k | 1 \leq k \leq c\}$

```
∀k, p_k ⇐ 1/c
∀k, ∀j, z_kj ⇐ 1/c
L_log ⇐ log L(p|g)
while TRUE do
    L^0_log ⇐ L_log
    // E-step
    for k = 1 to c do
        z_kj ⇐ p_k g_k(y_j)
    end for
    s ⇐ Σ_{k=1}^{c} z_kj
    for k = 1 to c do
        z_kj ⇐ z_kj / s // Normalize z_kj since it satisfies that Σ_{k=1}^{c} z_kj = 1
    end for
    // M-step
    p_k ⇐ Σ_{j=1}^{n} z_kj / n
    L_log ⇐ log L(p|g)
    if |L_log − L^0_log| ≤ ε then
        break
    end if
end while
return p
```

Table 6: Summary of the dataset

Name	Description
Collection time	Jan. 2013 to Nov. 2014
Number of air passengers	14,667,709
Number of records	99,308,782
Number of records per passenger per year	5

6. EXPERIMENTS

In the section, we present experimental results of the proposed model. First, we describe the dataset from an online travel agency and list several baselines. Then, a commonly used metric, log-likelihood, is computed to show the effectiveness of the proposed model. Finally, we offer some visualization results.

6.1 Dataset description

In our experiment, we use the dataset of an online travel agency. The system of the online travel agency allows passengers to book various kinds of flights from different airlines online. The dataset collects all booking records of domestic flights from January 2013 to November 2014. Each booking record contains the flight information and booking information, including *user id, flight number, booking time, airline, takeoff time, arrival time, departure port, destination port, price, price discount, seat class, passenger id, gender and age*. The details of the dataset are summarized in Table 6.

6.2 Experiment settings

Generally, the dataset is divided into two parts, from January to December 2013 and from January to November 2014, for training and evaluation, respectively. The set of active air passengers are chosen. In each run, we randomly choose 1,000 air passengers from the active passenger set. Log-likelihood is used as the metric for each air passenger and thus its average and standard deviation are collected. The listed results are generated by averaging the results of 20 runs.

6.3 Evaluated models

To validate the effectiveness of the proposed model, we evaluated its performance with the following counterparts. Suppose the number of components in the mixture model is denoted by c, and the bandwidth in kernel density estimation is denoted by h. All the evaluated models are constructed at the individual level.

- **GMM**: Gaussian mixture model is a commonly used model which employs the Gaussian distribution as the distribution function for each component. Number of components is chosen to be 2 or 3.
- **fKDE**: It is a fixed-bandwidth kernel density estimation for each air passenger. The bandwidth h is set according to the famous Silverman rule [13]. Gaussian function is used as the default kernel.
- **mix-KDE**: This is the proposed model. It is a mixture model whose individual component is expressed by kernel density estimation. The bandwidth of kernel density estimation is determined by the cross-validation

method. The default kernel is Gaussian kernel. In the experiment, we only test the mixture model with two components for simplicity, which is created with the passenger's own records and records of other passengers that share similar behavior with the passenger as a whole. Specifically, we set $\tau_1 = 0.3$ and randomly choose 1,000 passengers from all those passengers satisfying τ_1 as the set used for the second component.

6.4 Evaluation results

Log-likelihood is computed based on Equation (6), i.e., logarithm of likelihood, $log\mathcal{L}$. This metric shows how the proposed model fits the evaluation data.

Definition (**Likelihood**) Suppose the observed data samples are denoted by $\mathbb{D}_i$, the proposed model is $\mathcal{M}_i$, the likelihood is equal to the probability of those observed samples given the model. The equation is

$$\mathcal{L}(\mathcal{M}_i, \mathbb{D}_i) = P(\mathbb{D}_i|\mathcal{M}_i). \quad (6)$$

To evaluate the influence of impacting parameters, we vary impacting parameters and compare their results. Two models of GMM with different numbers of components ($c = 2$ and $c = 3$) are evaluated, which are denoted by GMM-2 and GMM-3, respectively. The bandwidth of kernel density estimation is chosen from 1, 1.5 or via cross-validation. Thus, the model is written as model name with bandwidth, for example, a mix-KDE method with a fixed bandwidth 1.5 is denoted by mix-KDE-1.5.

From evaluation results in Table 7, we can easily see that the proposed method (mix-KDE-cv) outperforms other methods both in average or standard deviation in terms of log-likelihood. GMM and fKDE have much larger standard deviations than that of the proposed method, which is probably because they can only fit some kinds of air passengers while fails to provide accurate description for other kinds of air passengers. In the table, we additionally offer the comparison results of GMM with different components. It is reported that more components doesn't necessarily improve the performance according to the average of log-likelihood. Furthermore, GMM with more components is more likely to suffer from over-fitting.

Table 7: Evaluation results of log-likelihood

Method	Average	Std.
GMM-2	-24.982	67.209
GMM-3	-45.016	40.919
fKDE-1.5	-11.364	9.439
mix-KDE-cv	**-7.030**	**2.277**

Table 8: Evaluation results of various bandwidths

Method	Average	Std.
fKDE-1	-13.178	21.103
fKDE-1.5	-11.364	9.439
mix-KDE-1	-8.423	1.241
mix-KDE-1.5	-9.523	**0.981**
mix-KDE-cv	**-7.030**	2.277

Table 9: Evaluation results of different kernels

Kernel	Average	Std.
Gaussian	**-7.030**	2.277
Exponential	-8.267	**1.220**
Tophat	-23.423	35.241
Epanechnikov	-20.756	33.743

In the next group of experiments, we compare the evaluation results of different models with different bandwidths. It is shown in Table 8 that the proposed data-driven method (mix-KDE-cv) improves the performance when compared to the predefined and fixed bandwidth determination schemes. For the fKDE model, larger bandwidth helps to improve the accuracy, while for the mix-KDE model larger bandwidth doesn't necessarily improve the performance in terms of average accuracy. It is reasonable that points that is far from the considering point may show different properties and it should be excluded when computing the density.

In Table 9, we compare the results of mix-KDE-cv with different kernels. The first two kernels, Gaussian and exponential, outperform the other two. The method mix-KDE-cv with tophat and epanechnikov kernels are different from the first two since those two only take account of points within the range of $[-h, h]$ for estimation. It tells that sometimes diverse air passengers are possible to provide useful information. The method with the epanechnikov kernel performs better than that with tophat kernel because the tophat kernel ignores different importances of points and sets identical weight for different points with $[-h, h]$.

6.5 Visualization

In the following, we offer the visualization of two groups of models from four different air passengers in Fig. 8 and Fig. 9. Other fields except number of advanced days (i.e., α) and hour type (i.e., $Hour$(takeoff time)) are fixed. The density of air travel choice behavior is shown in z-axis.

From the comparison in Fig. 8, we can conclude that both air passengers usually book their flights within 5 days in advance. Furthermore, the most frequent takeoff hour of them is 1 p.m. (i.e., 13 in the two pictures). However, the distributions of takeoff hours of these two passengers are different; passenger in the left are more likely to take a flights at night compared to his counterpart in the right.

In another groups of models shown in Fig. 9, the two air passengers are more probable to book flights much earlier, e.g., 10 days or more in advance, than those two in Fig. 8. That is to say, those air passengers in Fig. 9 usually make earlier travel plans. In another dimension of takeoff hours, those two passengers also have different distributions. The most frequent takeoff hour of them are 9 p.m. (i.e., 21 in the left figure) and 2 p.m. (i.e., 14 in the right figure), respectively.

7. RELATED WORK

In the section, we review the related work from the following three aspects and then point out their limitations.

7.1 High-level air travel behavior modeling

There exist a lot of studies, such as prediction of air travel demand [12, 14, 3], air traffic control [8], analysis of air travel

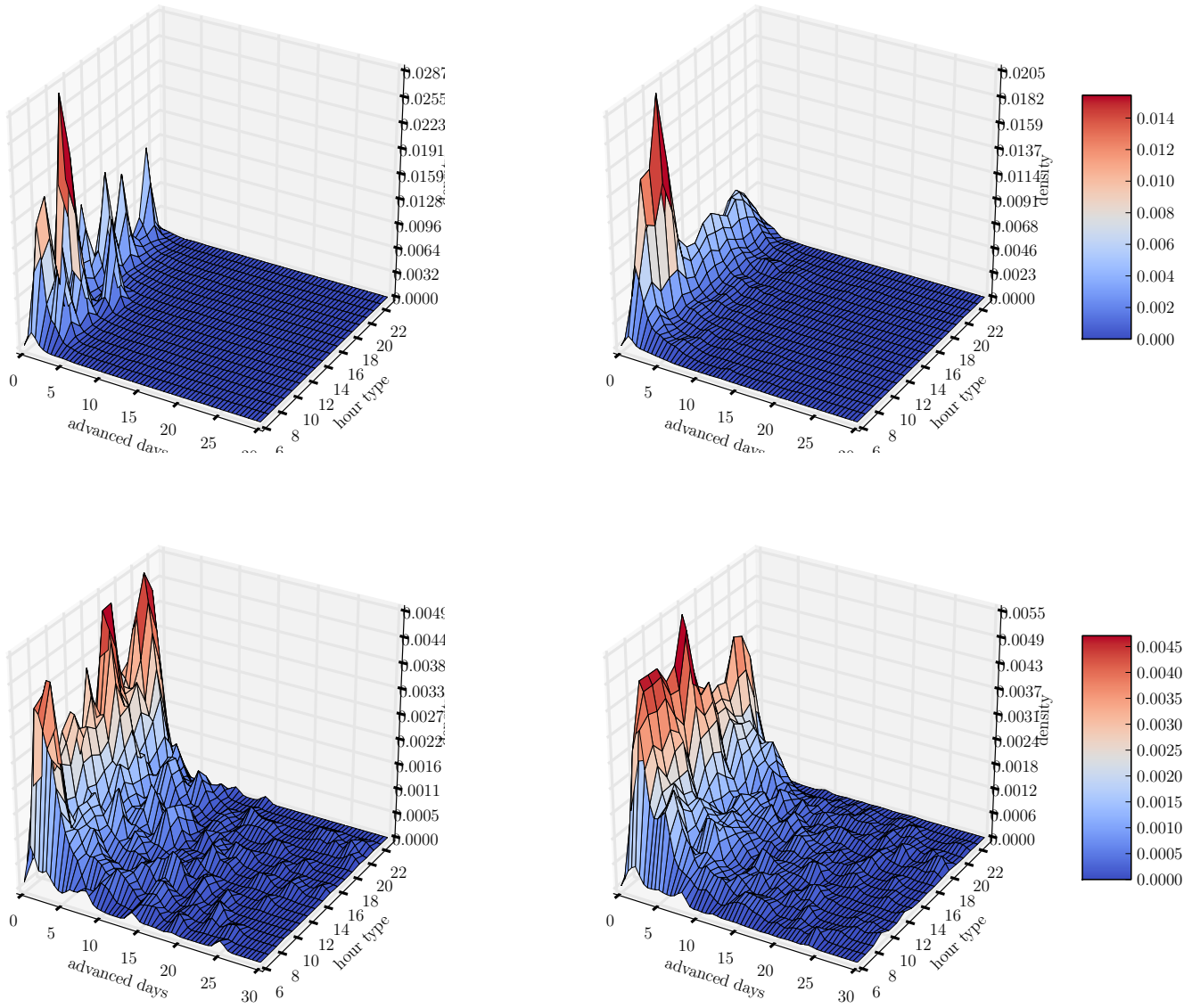

Figure 9: Visualization of models for another group of air passengers.

trends [11, 7], which are conducted to boost the air travel service market. These studies regard air passengers as a whole and analyze their behavior at an aggregated level. It is of great importance to airport operators or airlines, which require high-level statistical information to plan their space or investment for the aim of maximizing their profits. However, it is not suitable to employ an aggregated level air travel choice behavior model for online travel agencies which probably benefit from accurate recommendations towards a passenger.

7.2 Modeling with discrete choice models

In previous studies [2, 6, 5], revealed preference (RP) and stated preference (SP) data are widely used. These two kinds of data are generated by conducting surveys among air passengers. These datasets are extensively studied with the discrete choice models in [16, 4]. Modeling air passengers' behavior can be expressed as predicting individuals' choices from a given choice set. The discrete choice model is a commonly used mathematical model which describes and predicts one's choice from a finite set of alternatives. Each individual passenger, also called the decision maker, follows a decision rule to select a best item from all alternatives. Decisions rules show preference to different attributes associated with each alternative. In air travel choice behavior modeling problem, the service attributes include price, take-off time, seat and so on. Generally, the multinomial logit model and the mixed logit model are used to express the decision rules. However, survey data has its limitation since people sometimes don't understand their own behavior well enough or reluctant to provide real and objective comments about themselves.

7.3 Other studies of air travel behavior

Several other studies also consider understanding air passengers' behavior, such as [10, 9, 15]. In [10], the authors investigate the behavior of air passengers by highlighting several import factors that may affect air passengers' choice,

which was called price elasticity. The work is a kind of ticket pricing. In another work [9], the authors try to study how to make a trade-off between airports and airlines in multi-airport regions. Apart from price, non-price characteristics such as airport access time, flight frequency are found to closely affect choices of air passengers. Tam et al. consider the behavior of air passengers on choosing their ground access to airports [15]. Air passengers must reserve a period of time, which is called safety margin, for their trips to departure airports. The choice of their ground access modes is also regarded as a kind of behavior of air passengers.

8. CONCLUSION AND FUTURE WORK

In the paper, we have studied the problem of modeling the air travel choice behavior at the individual level with historical booking records. We apply the kernel density estimation for individual-level modeling. To tackle the data sparsity problem, a mix-KDE approach is proposed in order to exploit the broader pattern from a population of individuals sharing similar air travel choice behavior. The experimental results show the advantages and effectiveness of our proposed mix-KDE approach over GMM and simple kernel density estimation. Since the metric, likelihood, can be used a rank criteria for ranking or sorting algorithms, the proposed model could be easily used for recommender systems or personalized services.

There still exist a few issues which help to improve the proposed model. Bandwidth determination for multi-variate kernel density estimation is complicated. Some existing studies have been conducted to tackle the problem from different aspects. In our model, we did not take much consideration of the shape when choosing the bandwidth. In addition, an adaptive or local bandwidth determination method can be investigated. We leave them for future work.

Acknowledgments

This research is supported in part by 863 Program (No. 2015AA015303), 973 Program (No. 2014CB340303), NSFC (No. 61472254, 61170238 and 61420106010), STCSM (Grant No.14511107500 and 15DZ1100305), Research Grant for Young Faculty in Shenzhen Polytechnic (No. 601522K30015), SZSTI (No. JCYJ20160407160609492) and Singapore NRF (CREATE E2S2). This work is also supported by the Program for New Century Excellent Talents in University of China, the Program for Changjiang Young Scholars in University of China, and the Program for Shanghai Top Young Talents.

9. REFERENCES

[1] Statistical data of civic aviation administer of china at may 2016, 2016. http://www.caac.gov.cn/XXGK/XXGK/TJSJ/201607/P020160721351396626301.pdf.

[2] M. C. Bliemer and J. M. Rose. Experimental design influences on stated choice outputs: An empirical study in air travel choice. *Transportation Research Part A: Policy and Practice*, 45(1):63 – 79, 2011.

[3] M. Brons, E. Pels, P. Nijkamp, and P. Rietveld. Price elasticities of demand for passenger air travel: a meta-analysis. *Journal of Air Transport Management*, 8(3):165 – 175, 2002.

[4] E. Carrier. *Modeling the choice of an airline itinerary and fare product using booking and seat availability data*. PhD thesis, Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, 2008.

[5] S. Hess. Posterior analysis of random taste coefficients in air travel behaviour modelling. *Journal of Air Transport Management*, 13(4):203 – 212, 2007.

[6] S. Hess. Treatment of reference alternatives in stated choice surveys for air travel choice behaviour. *Journal of Air Transport Management*, 14(5):275 – 279, 2008.

[7] S. Hess and T. Adler. An analysis of trends in air travel behaviour using four related sp datasets collected between 2000 and 2005. *Journal of Air Transport Management*, 17(4):244 – 248, 2011.

[8] L. Hu, F. Sun, H. Liu, and H. Xu. Flight behavior recognizing in terminal area based on support vector machine. In *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*, 2010.

[9] J. Ishii, S. Jun, and K. V. Dender. Air travel choices in multi-airport markets. *Journal of Urban Economics*, 65(2):216 – 227, 2009.

[10] M. Kouhpaei. Airfare price elasticity over non-business passengers. In *Information and Financial Engineering (ICIFE), 2010 2nd IEEE International Conference on*, 2010.

[11] T. Kuhnimhof, R. Buehler, M. Wirtz, and D. Kalinowska. Travel trends among young adults in germany: increasing multimodality and declining car use for men. *Journal of Transport Geography*, 24:443 – 450, 2012.

[12] T. Li, H. Baik, and T. Spencer. An optimization model to estimate the air travel demand for the united states. In *Integrated Communications, Navigation and Surveillance Conference (ICNS), 2014*, 2014.

[13] B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[14] E. Suryani, S.-Y. Chou, and C.-H. Chen. Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework. *Expert Systems with Applications*, 37(3):2324 – 2339, 2010.

[15] M. L. Tam, W. H. Lam, and H. P. Lo. Modeling air passenger travel behavior on airport ground access mode choices. *Transportmetrica*, 4(2):135–153, 2008.

[16] V. Warburg. *Modeling air travel behavior*. PhD thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark, 2005.

[17] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.