Survival Analysis for Delay Prediction Based on Queue Models

Ohnmar Aung
Faculty of Computing,
University of Computer Studies, Taunggyi
Ohnmaraung.fc@ucstgi.edu.mm

Nang Win Phyu Phyu Naing
Information Technology Supporting and Maintence,
University of Computer Studies, Taunggyi
winphyuphyunaing@ucstgi.edu.mm

Abstract

In recent years, queue models have been many interesting problem and popular topic that have been applied to the various research areas in real world applications. Queue Models 'greatest success in the real-world application areas have been in communications and data networking. In this study, a delay predictor for waiting customers outside restaurant is developed based on queue models such as Enter Service Rate of Last Customer and Head of Waiting Line are used to estimate the condition of waiting in the restaurant and the effect on the out of waiting status. For this work, Cox Model among survival analysis methods is used to calculate the service rate and delay time as Status of Customer Service and Gradient Boosting Regression Tree (GBRT) is used. The experimental results of this system show that the well work of this predictor and this proposed design is applied for real-world applications.

Keywords: Queue Model, Cox Model, Gradient Boosting Regression Tree, Status of Service, Last Customer Enter Service, Survival Analysis;

1. Introduction

A leading subject of studies in operation research, queue models are the mathematical studies of queue and waiting lines. Optimal control of queue systems extends the idea of optimization to dynamically varying service rates, arrival rate and other variables. Differences of simulation methodologies, queue models need a little dataset and sample related results to formulate for prediction various measure of performance such as mean value for delay or probability of waiting time than a predefined amount of time before being served. In recent application, the delay prediction is common problem. For popular restaurants, customers always wait to get the service and to understand how many customers a head of them.

The delay prediction system can support these restaurants to manage time if they can be known how many customers have in the queue to wait the service. Since 2009, the queue models have been applied to represent the telephone exchange, various queue models are widely used for different senses in real scenarios such as scheduling system [1]. In recent decades, prediction of delay has been a main article in academic field in real world as an effective predictor for delay that can improve experiences of user mostly.

A general scene for prediction on delay is the traffic system management. Mukherjee developed queue models for prediction the status on delay time of flight by condition of bad weather and delay schedule of traffic [2]. Nakibly implemented telephone service system based on queue models to predict the delay [3]. The combination of queue models with the processes of mining in customer service model is developed by Senderovich [4].

The main factors for user experience in call center are identified by Finberg and then the duration of delay that influences the independent rate in significantly is proved that according to the experimental results [5]. The on Radial Basis Function Neural Network is created for prediction the packet end-to-end delay for mobile ad hoc networking [6]. An enhanced Survival Analysis Model was developed by Jian Wan that named expected model for the prediction of the probabilities with user buying a recommend product at a provided time [7]. The Cox Model the probabilities rearrested after they were observed among times [8].

In this proposed system, the queue model and Survival Analysis are used to extract the features that used as input for GBRT to achieve the better accuracy on delay prediction. Especially, the different queue models based on customer service and waiting line are designed by this proposed system that used Survival Analysis for the service rate estimating and condition in the restaurant to evaluate the delay rate for restaurant. The using of queue data from real application is implemented to make this work and the delay predictor is created to work well in real world application.

2. Background Theory

To predict the lengths of queue and waiting time, Queue Model is very popular mathematical study. Because the results of queue model are applied when decision making in business related to the required resources to give a service, Queue Model is mostly developed as a branch of operation research.

Queue Model is a branch of Operation Research that can help user for making the business decision on how to build the efficient and cost-effective workflow system. Arrival rate, service time and number of servers in the system are very important measurement variables for Queue Model. The input assigns state, discipline of queue service, queue models are involved in the queue system as critical features. As input assign state, customers may be infinite numbers or finite numbers, customer can be arrived group by group or only one arrival, the arrival rate can be independent, and the inter-arrival rate cannot be dependent to time. The queue model discipline are as customers queue for service, customers has no servers available, queue is not enough for customers or they have already in the waiting queue length for a long time. The First In First Out, Last In First Out, Priority in order to service and Random in order to service are the service processes of Queue Model. The Single Server, Parallel Multi Servers and Hybrid or Series Multi Servers are the kind of Queue Server. A delay predictor can be divided into delay predictor based on queue length and delay predictor based on delay. Delay predictor can be predict based on number of queue length of waiting customers and delay history predictor can be calculated on knowledge of customers' waiting history.

2.1. GI/ M/ s Model

The considering a stationary balking sequence in GI/M/s queue model is very important part. The average rate of Finite horizon achieved criterion is among all policies stationary that the optimal results are not random for the rules of control limit: join if and only if the queue size is smaller than some define number.

The queue model formalization can be defined as $\alpha/\delta/\theta/\beta/\gamma$. The interval time for arrival rate distribution of customers can be represented as α. In this formulation, δ express customer service distribution rate, θ denotes count of servers used in system, β means the system capacity and γ represents the discipline of service. For general model γ is FIFO and β is restricted [9]. The foundation of *GI/M/s* model can be defined as following descriptions:

- *GI*: Distribution of Arrival rate.
- λ: Average customers arrival
- *1/*λ: Arrival rate between two customers
- *M*: Service rate in the system
- μ: Average number of single services
- $1/\mu$: Average service
- *s*: Parallel servers' number.

The mean of the GI/M/s is a queue model with parallel servers s. The positive random variables are identically distributed with cumulative distribution function and the inter arrival time is independent. The exponential random variables are identically distributed with mean vale 1 using independent average customer service rate for each customer. To calculate the service rate of the system, the following equation can be used:

$$Service\ Rate = \frac{1}{\mu}$$

And then the delay time of the system can be calculated by using the following equation:

$$Delay\ time = \frac{1}{(\mu - \lambda)}$$

The example calculations for service rate and delay time are shown in following:

Let λ = 10 customers per hour
μ = 5 customers per hour

$$\begin{aligned} Service\ Rate &= \frac{1}{\mu} \\ &= \frac{1}{5}\ hours \\ &= \frac{1}{5} \times 60\ minutes \\ &= 12\ minutes \end{aligned}$$

Service Rate for each customer is 12 minutes.

$$\begin{aligned} Delay\ time &= |\frac{1}{(\mu-\lambda)}| \\ &= \left|\frac{1}{(5-10)}\right| \\ &= \frac{1}{5} \times 60\ minutes \\ &= 12\ minutes \end{aligned}$$

Delay time is considered on the range of duration time, negative value is not being considered. According to the arrival rate and number of service rate, the delay time can have 12 minutes for that system. So, the service rate and delay time depend on the arrival rate of the system and number of customers in the system.

2.2. Survival Analysis

After certain time, to evaluate the events occurrence probabilities, the Survival Analysis was developed [10]. The survival function and hazard function are two fundamental functions in Survival Analysis.

The individual lifespan can be defined as *T*, survival function can be represented as *S(t)* for the death probability that has not been caused until time *t.* The following equation is used to calculate the Survival function:

$$S\ (t) = P\ (T>t),\ 0 \leq S\ (t) \leq 1$$

S (t) is a decrease function of *t.* In Survival function, the point *x-axis* represents for rate of survival and the point *y-axis* represent for the probabilities of survival.

The individual probability who survived at time t is described in Hazard function *h(t) but* fall down in t + δt:

$$h(t) = \lim_{\delta t \to 0} P(t \leq T \leq t + \delta t \mid T > t)/\delta t$$

Survival analysis models are parameters represented method (parametric), semi-parametric method and non-parametric method. Parametric is used for the survival time according to a specific distribution. This method is not in common use as the specific distribution is usually unknown. Kaplan Meier, non-parametric method is applied to keep statistics on dataset and to build a model evaluation of assays how the factors impact the survival time. Cox model, Semi-parametric method does not need to know the distribution of survival time and to build a model with assays the influence factors.

2.3. Cox Model

In 1972, using Cox, Cox proportional Hazard Model was developed to analysis on the survival time affected by covariates [11]. A method for investigating the effect of several variables is Cox model upon the time a specified take with happen events. A statistical technique, Cox Model is will-recognized for exploring the relationship of the survival and various explanatory variables. One of the most important methods for modelling survival analysis data is Cox hazard model. In a Cox hazards model, the rates of hazards are the effective measurement, given that the participants have survived to a predefined time. The Cox model hazard function is defined as:

$$h(t) = h_0(t)\exp\,(b_1x_1 + \cdots + \ b_Nx_N)$$

h(t) defines the individual using *N* variable at the moment *t*, the model evaluates the individual probabilities to survive at *t* but dead at the next clock. $h_0(t)$ is the rate of hazard baseline which is same as $h_0(t,0)$ and the coefficient of variables x id b. The vice versa and the risk of hazard are increased by using the positive coefficients that denoted corresponding covariates. The rate of individual hazard is proportionated the rate of baseline hazard at the random time t:

$$h(t)/h_0(t) = \exp\,(b_1x_1 + \cdots + \ b_Nx_N)$$

For suitable use of Cox model hazards regression, there are many various important assumptions that including the following descriptions:

- Independence survival times between different individuals
- Improved multiple relationship between the hazard rates and predictors and the hazard
- Constant ratio of hazard rate over time

2.4. Gradient Boosting of Regression Trees (GBRT)

Gradient Boosting of Regression Trees (GBRT) was firstly developed by Friedman, for estimating the car-following model. The key goal in GBRT algorithm is to fit a regression tree to the difference between the observed response and the aggregated prediction of all learners grown previously. As indicated by the name of the algorithm, the decision trees algorithm with a fixed size is chosen as the weak leaner. The core idea in ensemble methods is to combine multiple base learners, which are usually called weak learners, in order to improve the accuracy and robustness of the final model. Two major questions need to be answered in ensembles methods: what type of weak learner should be used for the training process? How to train and fuse weak learners to yield the final model? Accordingly, ensemble algorithms fall into the two main families: averaging and boosting methods. Each family follows a distinct procedure towards training and combining individual base learners. In the averaging methods, base learners are first built independently on randomly selected training instances, and then are averaged to generate the final model.

3. Proposed Algorithmic Method

To apply the various improved queue model's output and important features for GBRT model as the input and to achieve the better performance of delay prediction is the main idea of delay predictor.

3.1. Factual Features

The following factual features represent for the basis characteristic of queue instance.

- Seat type: Different size of table means different seat type. The waiting time more cost because the oversize tables are usually rare than normal size tables.
- Weekday: During Weekend, there always is more outside waiting customers but the inside conditions of restaurants are not same among weekday.
- Hour: At the peak hour, delay rates are more suitable in studying and more similarly to be effective development.
- Queue length: Queue length is the count of customers at the waiting queue for same type of seat lead to the current customers.

3.2. Delay History based Queue Model

Nowadays, the history-based delay predictors are easy to develop and built [12]. The predictor for delay rate was developed by M. Armony to evaluate the equivalent performance of system [13]. To begin with, defined $U_n$as inter-arrival rate as unit time and the arrival state is defined by:

$$A(t) = max\{n \geq 0 : U_1 + \cdots + U_n \leq t\}, t \geq 0$$

The last customers delay rate entered the service is defined as standard *w* and the arrival rate (unit in time) of new customer is defined as *t.*

$$W_{LES}(w,t) = w$$

To get more information and to evaluate the length of queue *y* the delay rate of customer to get the service. The following equation can be represented for delay predictor:

$$W_{LES}(w,d) = \sum_{i=1}^{A(w+d)+1} (S_i/s)$$

In this equation, w represents the experienced delay of last customer to enter service, d defines the exit time since the customer enter service. t_a is the time for arrival customer and t_e is the time for the last customer entered service before ta. Therefore, $d= t_a - t_e$. If t_e is simultaneously to the last service completion prior to ta, in other words, there is at least one customer left in queue at t_e. The length of queue at t_a distributed as $(w + d)$. Si is the rate of service for customers and the number of servers in the system is s. In this system, the previous approximately observed data is s and S_i is evaluated by Cox model. Si, the proportion surviving (or remaining event free) past interval i; this is sometimes called the cumulative survival probability and it is computed as follows: First, the proportion of participants surviving past time 0 (the starting time) is defined as S_0 = 1 (all participants alive or event free at time zero or study start). The proportion surviving past each subsequent interval is computed using principles of conditional probability introduced in the module on Probability. Specifically, the probability that a participant survives past interval 1 is S_1 = p1. The probability that a participant survives past interval 2 means that they had to survive past interval 1 and through interval 2: S_2 = P(survive past interval 2) = P(survive through interval 2)*P(survive past interval 1), or S_2 = p2*S_1. In general, S_{i+1} = pi+1*S_i.

In delay predictor, w denotes the already waiting rate of the customers at the head of the line:

$$W_{HOL}(w) = \sum_{i=1}^{A(w)+2} (S_i/s)$$

According to the service inside condition of restaurant, the queue model is used for estimation of delay state on the current customer's rate. The main idea of model is to evaluate the waiting customers' rate as queue and the number of customers to get the service into the restaurant at the end of queue length. Array A defines the number of active eating table of the restaurant. Applying the Cox model for each item in A to evaluate the time before t finished as array R. In real scene, the length of high waiting rate always provides the high left customers rate, so the β is used for evaluate the queue length for real scene from the left customers. The following equation can be represented for the SOS delay predictor:

$$W_{SOS}(w) = R[\beta * queue_{length}]$$

The different kinds of delay predictors are use among Cox model to evaluate the unknown customers' service rate (unit in time). The length of queue describes the waiting status of customers out of the restaurants and the length of order describes the status of the restaurant's service The logs of order are divided into sets by various restaurants and independent hour starring time. According to the service rate overcame by starting service rate. This model is trained for each period and numbers of peoples.

4. Experimental Analysis

4.1.Data Set

In this system, 100 restaurants are collected and used for the service system using queue model August 2018, only 6 restaurants of them to get over 600 records of queue. This work based on 6 restaurants that have logs of queue 12,000 records and logs of service 15,000 records at total service rate (units in time). Mean value absolute error and median value absolute errors are calculated for performance evaluation result. In real-world application, these predictors are useful product to support the customers, the demand of the system is to control the deviations in 10 minutes.

4.2. Performance Analysis

In this work, the important feature function is used to calculate relative features of the system. The mean value and 3-folds cross validation are used to evaluate the important feature and to generate the performance of this system. The absolute mean and median error are calculated for all average value of 3-folds and the results of 3-folds are described in table 1.

Table 1. Analysis of Mean and Median Value Error

Restaurants	Mean value Error	Median value Error
A	11.12	7.25
B	12.01	7.25
C	21.1	7.89
D	11.99	8.12
E	16.98	11.9
F	8.25	5.35

Table 2. Distribution of Absolute Error for Restaurants between 10 and 20 minutes

Restaurants	within 10 minutes	Between 10 and 20 minutes	Total Range on Records
A	2012	875	3245
B	52	19	99
C	204	95	357
D	72	32	125
E	72	48	182
F	146	39	203

In table 2, the absolute distribution errors are also described as the experimental results. If the absolute error rate between prediction and actual delay during ten minutes, it can help for waiting customers in queue. According to the experimental results, the delay predictor can work well on given dataset but some counterexamples are still as the results.

The case of deep deviation, the one by one record is collected and the reason can be listed to three categories:

- General scenes: In most case, an unsuitable result is given by the predictor compare with the actual delay.
- Long delay: Long delay is a time that longer than 60 minutes of actual delay time. Because of this condition, the delay of predictor is usually less than real delay and 20 minutes can exceed than absolute error.
- Short delay: The short delay is smaller than five minutes of actual delay. According to this state, the delay of predictor usually larger than the real delay and the 20 minutes of absolute error rate can be exceeded at some times.

5. Conclusion

Queue model is very effective and complex system that involves different aspects. In this system, the reactions of customer's feedback for delay are not considered but other ideas of queue system are considered. Because of the experimental result, if the queue length is too high, the customer commonly can be given up. In this condition, the customers can be missed for their queue and this reaction will be affected to the other customers in the queue state. Using this queue model, the management of restaurant can be solved the problem about the waiting customers and can be improved their service rate for customers. According to the analysis of this system, Survival analysis and Queue Model can be performed and can be well combined with GBRT.

References

[1] A. Daniel, and et al., *"Performance by Design: Computer Capacity Planning by Example,"* Prentice Hall, 2004.

[2] A. Mukherjee, and et al., *"Predicting Ground Delay Program At An Airport Based On Meteorological Conditions,"* 14th AIAA, 2014.

[3] E. Nakibly, *"Predicting Waiting Times in Telephone Service Systems,"* TIIT, 2002.

[4] A.Senderovich, and et al., *"Queue Mining Predicting Delays in Service Processes,"* TIIT, 2014.

[5] R. AFeinberg, and et al., *"Operational determinants of caller satisfaction in the call center,"* IJSIM, 2013, vol. 11, pp. 131- 141.

[6] J. P. Singh, and et al., *"Delay Prediction in Mobile Ad Hoc Network using Artificial Neural Network,"* PT, 2012, vol. 4 (11), pp. 201-206.

[7] J. Wang and Y. Zhang, *"Opportunity Models for E-commerce Recommendation: Right Product, Right Time,"* IASCRDIR, 2013.

[8] J. Fox and S. Weisberg, *"Cox Proportional-Hazards Regression for Survival Data in R,"* R and S-PLUS, 2002.

[9] G. David, *"Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain,"* AMS, 1953, pp. 338-354.

[10] S. J. Richards, "*A handbook of parametric survival models for actuarial use,"* SAJ, 2012, vol. (4), pp, 233-257.

[11] R. David, *"Regression Models and Life-Tables,"* JRSS, 1972, pp. 187-220.

[12] A. Mandelbaum and N. Shimkin, *"A model for rational abandonments from invisible queues,"* Queueing Systems, 2000, vol. (36), pp. 141-173.

[13] M. Armony, and et al., *"The Impact of Delay Announcements in Many-Server Queues with Abandonmentible queues",* Operations Research, vol. (57), pp. 66-81,2009.