On Predicting a Call Center's Workload: A Discretization-Based Approach

Luis Moreira-Matias[1,2,3], Rafael Nunes[3], Michel Ferreira[1,5],
João Mendes-Moreira[2,3], and João Gama[2,4]

[1] Instituto de Telecomunicações, 4200-465 Porto, Portugal
[2] LIAAD-INESC TEC, 4200-465 Porto, Portugal
[3] FEUP, U. Porto, 4200-465 Porto, Portugal
[4] Faculdade de Economia, U. Porto 4200-465 Porto, Portugal
[5] DCC-FCUP, U. Porto, 4169-007 Porto, Portugal
{luis.m.matias,joao.mendes.moreira,jgama}@inescporto.pt,
{rafael.nunes,michel}@dcc.fc.up.pt

Abstract. Agent scheduling in call centers is a major management problem as the optimal ratio between service quality and costs is hardly achieved. In the literature, regression and time series analysis methods have been used to address this problem by predicting the future arrival counts. In this paper, we propose to discretize these target variables into finite intervals. By reducing its domain length, the goal is to accurately mine the demand peaks as these are the main cause for abandoned calls. This was done by employing multi-class classification. This approach was tested on a real-world dataset acquired through a taxi dispatching call center. The results demonstrate that this framework can accurately reduce the number of abandoned calls, while maintaining a reasonable staff-based cost.

Keywords: call centers, arrival forecasting, agent scheduling, discretization, multi-class classification.

1 Introduction

Staffing is one of the major problems in call center management. This paper focuses on predicting workload in call centers in order to improve staff scheduling. By using these predictions, it is possible to formulate the scheduling problem as an optimization problem. The goal is to minimize the number of abandoned calls. This problem can be simply solved by using a heuristic method of interest.

Workload estimation focuses on predicting demand. In the literature, scheduling is formulated according to a **point-wise** prediction of the quantitative target variable. To address this problem, it is possible to divide the predictive models into two types: (a) time series analysis and (b) regression. The first one (a) typically relies on assuming homogeneous or time-varying Poisson processes to *feed* Holt-Winters smoothing models or ARIMA-based models [1,2]. The second type establishes a relationship between the number of arrivals and other explanatory variables such as the day type [3,4].

T. Andreasen et al. (Eds.): ISMIS 2014, LNAI 8502, pp. 548–553, 2014.

Type-a approaches typically assume the future number of arrivals as a linear combination of their historical values. By dealing with large time horizons (days), these models *lose* one of their best assets: the ability to react to *bursty* events [5]. Type-b models are able to establish more complex relationships (e.g. non-linear relationship in Artificial Neural Networks (ANN)). However, many of these models aim at minimizing the root mean squared error (RMSE) between the predicted and the actual arrivals, discarding demand peaks (i.e. outliers). Consequently, these extreme events represent the highest ratio of *abandoned* calls. The compromise between (a) *understaffing* to maintain low costs by losing *some* service demand and (b) *overstaffing* to minimize abandoned calls is hardly done by assuming constant or periodic arrival rates. Consequently, the workload forecasting problem may not be adequately addressed by such methods.

Discretizing continuous variables is a relevant building block of many machine learning algorithms (for instance, C4.5). Hereby, this paper proposes a *local discretization* technique to address this limitation. In this applicational framework, the basic idea consists of determining the number of agents required to meet a demand expressed in **equal-width** intervals. The interval width corresponds to the **expected agent capacity**, which is assumed to be *constant* over time and along the different workers. By dividing the arrival counts into equal-width intervals, the goal is to accomplish two distinct goals: (1) reducing the search space for supervised learning methods and (2) adding the classification methods to the current *pool* of problem solvers. This is a step forward towards solving the problem. The approach that is closest to the one presented here is the one proposed by Shen *et al.* [6]. The authors perform singular value decomposition to reduce the dimensions of the independent variables. However, they still formulate the workload prediction as a numerical forecasting problem.

A small call center with 13 agents running in the city of Porto, Portugal, was chosen as the case study for this paper. Its scheduling is still performed empirically. The results highlight the method's contribution to this problem as it outperforms the numerical prediction methods on the proposed dataset.

This paper is organized as follows: Section 2 describes our methodology to overcome the agent scheduling problem. Section 3 introduces the real-world scenario addressed in this study. The experimental setup and the results obtained are described in Section 4. Finally, conclusions are drawn in the last section.

2 Methodology

Let A_t denote the a stochastic process describing the number of arrivals per period of time. Let $N = \{n_i | n_i \in \mathbb{N} \wedge n_i \leq \Gamma\}$ denote the domain of the arrival counts A_t, where Γ is the maximum admissible value of arrivals per unit of time. This work explores two distinct predictive approaches: (1) one where $A_t \in N$ is used and another (2) where a value *interval* $\pi_i = [b_i, b_{i+1}) \in \Pi$ for A_t is predicted, such that $b_i \leq A_t < b_{i+1}$. Let $\Pi = \{\pi_i | \pi_i = [b_i, b_{i+1}) : b_{i+1} - b_i = b_i - b_{i-1}, \forall b_i \in N\}$ stand for the interval set and $\delta = b_{i+1} - b_i$ define the corresponding **width** interval. By reducing the target variable dimension, the goal is to enhance the detection of future workload peaks.

Let $X = \{X_1, ..., X_\rho\}$ denote the set of ρ *attributes*. Let $x = \{x_1, ..., x_\rho\}$ be a set of attribute values where $x_i \in X_i$. The goal is to infer the function $\hat{f}(x) \sim f(x) \in N : f(x) = A_t, \forall x$. This induction is data driven as it uses a training set T (i.e. a data set where each sample is a pair (x, A_t)) to compute the approximation. The learners usually perform multiple scans over T to iteratively update its models. This cycle stops when it finds a minimal value for a function which establishes the differences between $f(x)$ and $\hat{f}(x)$. This function is known as *objective function*. However, many learners tend to present rough approximations to bursty peak values by *smoothing* their models (e.g. Linear Least Squares). This effect is a major problem in this context. To overcome it, we propose to **discretize** the target variable. Let us redefine the problem as

$$\hat{f}(x) \sim f(x) \in \Pi : f(x) = \pi_i \Rightarrow A_t \in \pi_i, \forall x \tag{1}$$

Staff scheduling is a *bounded* problem, although the arrival prediction is not. Each call center is constrained to the number of workstations available (denoted as w), and (b) to the number of agents available, i.e. O_m. Let $C_m \in N$ be the maximum workload supported without calls abandoned by each agent per unit of time, and l_t be the number of abandoned calls during the same period. Let $O_w = \min(O_m, w)$ be the maximum number of agents operating per unit of time. The maximum workload supported by the call center without abandoned calls is defined as $\Psi = O_w \times C_m$. Consequently, it is not that relevant to predict whether $A_t > \Psi$ or $A_t \gg \Psi$. Discretizing attributes is a well-known preprocessing technique in machine learning. This paper proposes to discretize the independent variable A to reduce its domain length. In this context, the interval width can be defined as $\delta = C_m$. Therefore, it is possible to redefine the domain of $f(x)$ as $|\Pi| = O_w + 1$. It corresponds to O_w equal-width intervals plus an extra interval where $A_t \in \pi_i = [O_w \times C_m, \Gamma) : l_t = A_t - \Psi > 0$. By predicting an interval instead of an exact value, it is possible to formulate this problem as a *multiclass classification* problem where the label is the interval where A_t will fall into.

Estimating agent capacity C_o in a Call Center is a challenging problem. For the sake of simplicity, it is assumed that this is time-constant for each prediction. Let $Z \subseteq T$ be the dataset describing historical demand peaks defined as

$$Z = \left\{ z_i = \frac{A_t - l_t}{O_t} | \frac{l_t}{A_t} \geq \alpha > 0, \forall t \right\} \tag{2}$$

where O_t is the number of agents operating in t and α is a user-defined parameter to consider a past arrival count as a bursty event. C_m can be obtained as the **median** number of calls answered by each agent during a bursty event (i.e. $\tilde{Z}$).

The predicted π_t stands for the *desired* workload at time period t. Let H_i denote the number of shifts assigned to an agent i and Ω be its maximum. Let $b_{i+1,t}$ denote the upper-bound value of the predicted arrival count interval π_t. $O(t)$ can be computed as follows

$$\underset{O(t) > 0}{\arg\min}\ b_{i+1,t} - (O(t) \times C_m),\ s.t.:\ H_i \leq \Omega, \forall\ i, t \tag{3}$$

3 Case Study

The case study presented here is based on a taxi dispatching center in Porto, Portugal. The center distributes calls to a fleet of 441 vehicles and employs 13 agents. However, only $O_m = 11$ are available for scheduling due to the existing labor regulations. Their assignments are still performed on a weekly basis. Each agent can only have a maximum number of $\Omega = 5$ shifts assigned per week. However, the number of workstations available is $w = 4$. The call arrivals between June and October 2013 were used as test-bed dataset.

4 Experiments and Results

The first 17 weeks were used as a training set, while the last four were considered a test set. The training set for each week consisted of all the arrival count data available until this point in time. The α value was set as 0.15. An interval resolution of 30 minutes was considered in the experiments (as in [7]).

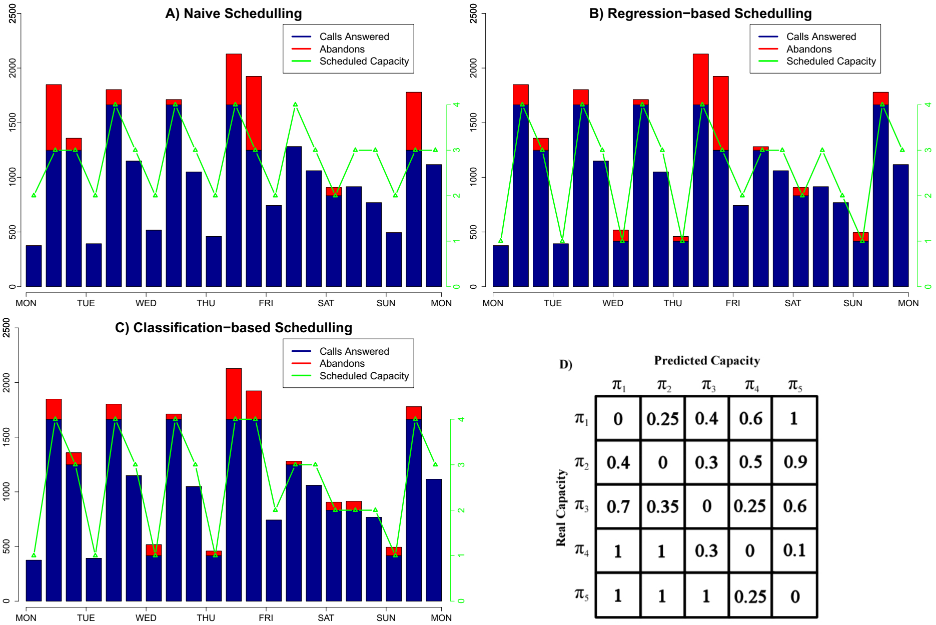
Two distinct approaches were followed in the experiments: (a) one where the arrival counts were predicted as an exact number and another (b) where they were predicted as an interval. In (a), the Holt-Winters smoothing was employed, as well as the k-Nearest Neighbors (kNN), the Random Forests (RF) and the Projection Pursuit Regression (PPR). In (b), the classification methods employed were the NB, RIPPER, ANN and RF. The classification methods were used with their default parameters. A sensitivity analysis was conducted on the parameter setting of each one of the regression methods employed using a simplified version of the sequential Monte Carlo method [8].

The scheduling problem formulated in the eq. 3 was solved using a Genetic Algorithm. The scheduling performed by the company was compared with the scheduling procedures generated by using the methods in (a,b). Since the output of (a) is an exact count, the π_t was replaced with the exact predicted count A_t.

In (a), the Mean Absolute Error (MAE) and the RMSE were employed as evaluation metrics. In (b), an accuracy-based metric and a user-defined cost-sensitivity matrix (it is more important to predict a peak than a normal arrival count) were used. That matrix expresses the cost of every misclassification case. This is shown in Fig. 1-D. To evaluate the scheduling quality, the number of expected abandoned called was computed in the different schedules proposed.

Table 1 presents an evaluation of the proposed algorithms that were used to predict workload. The results were evaluated based on the estimated capacity of each agent (i.e. C_m) by considering as "lost calls" all the calls besides the scheduled workload. The regression/classification method with the lowest averaged error in each week was used to perform scheduling in the following week. The sensitivity analysis results were used to select the methods for week 1. The scheduling results are provided in Table 2 and in Fig. 1.

Table 1 denotes an excellent performance using both regression/numerical and classification methods. However, the classification approach is almost as flawless as its accuracy $\simeq 1$. The evaluation performed in Table 2 and in Fig. 1 contains



	π_1	π_2	π_3	π_4	π_5
π_1	0	0.25	0.4	0.6	1
π_2	0.4	0	0.3	0.5	0.9
π_3	0.7	0.35	0	0.25	0.6
π_4	1	1	0.3	0	0.1
π_5	1	1	1	0.25	0

Fig. 1. Detailed Results of week 3 using different methods to estimate the workload (A,B,C) and the error cost matrix employed for the workload misclassification (D). Note the classification's refinement on Thursday evening.

Table 1. Results for the prediction task using both Regression/Numerical and Classification algorithms. The metric values were averaged for all weeks.

Numerical Prediction	**Holt-Winters Smoothing**	**Random Forests**	**KNN**	**PPR**
RMSE	17.93	19.31	21.31	22.38
MAE	11.52	13.32	13.55	16.14
Error Cost Matrix	0.025	0.029	0.027	0.034
Multi-Class Prediction	**NB**	**Random Forests**	**RIPPER**	**ANN**
Accuracy	0.938	1.000	1.000	0.805
Error Cost Matrix	0.020	0.000	0.000	0.096

Table 2. Abandoned calls using the different agent schedulings

	Week1	**Week2**	**Week3**	**Week4**	**Total**
Naive Method	5708	4819	4335	4541	19403
Numerical Prediction	4362	3449	4153	**3951**	15915
Multi-Class Prediction	**4306**	**3421**	**3981**	3961	**15669**

an error since the abandoned calls expressed are merely an expected value based on the scheduled capacity. However, Fig. 1 uncovers the limitations of the naive approach as it clearly overstaffs the night shifts. Not surprisingly, the greatest advantage of employing classification methods in this problem is their capacity to uncover **demand peaks**, as expressed in Fig. 1. These results demonstrate that the **interval-based classification** approach should be regarded as a reliable solution to this problem.

5 Final Remarks

This paper proposes a discretization-based framework to address the workload prediction with the calls made to the call center. The goal with this framework is to accurately predict demand peaks in order to optimize the use of resources. The results obtained show that this problem can be successfully addressed as an interval-based multi-class problem. The authors' goal was to use these findings as proof of concept to open new research lines on this topic.

Acknowledgments. This work was supported by the Project "I-City for Future Mobility" (NORTE-07-0124- FEDER-000064).

References

1. Avramidis, A.N., Deslauriers, A., L'Ecuyer, P.: Modeling daily arrivals to a telephone call center. Management Science 50(7), 896–908 (2004)
2. Taylor, J.W., Snyder, R.D.: Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega 40(6), 748–757 (2012)
3. Weinberg, J., Brown, L.D., Stroud, J.R.: Bayesian forecasting of an inhomogeneous poisson process with applications to call center data. Journal of the American Statistical Association 102(480), 1185–1198 (2007)
4. Millán-Ruiz, D., Hidalgo, J.I.: Forecasting call centre arrivals. Journal of Forecasting 32(7), 628–638 (2013)
5. Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., Damas, L.: Predicting taxi-passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems 14(3), 1393–1402 (2013)
6. Shen, H., Huang, J.Z.: Forecasting time series of inhomogeneous poisson processes with application to call center workforce management. The Annals of Applied Statistics, 601–623 (2008)
7. Aldor-Noiman, S., Feigin, P.D., Mandelbaum, A.: Workload forecasting for a call center: Methodology and a case study. The Annals of Applied Statistics, 1403–1447 (2009)
8. Cappé, O., Godsill, S., Moulines, E.: An overview of existing methods and recent advances in sequential monte carlo. Proceedings of the IEEE 95(5), 899–924 (2007)