K Nearest Sequence Method and Its Application to Churn Prediction

Dymitr Ruta, Detlef Nauck, and Ben Azvine

British Telecom (BT) Group, Chief Technology Office
Adastral Park, Orion MLB 1, PP12, Ipswich IP53RE, UK
dymitr.ruta@bt.com

Abstract. In telecom industry high installation and marketing costs make it between six to ten times more expensive to acquire a new customer than it is to retain the existing one. Prediction and prevention of customer churn is therefore a key priority for industrial research. While all the motives of customer decision to churn are highly uncertain there is lots of related temporal data sequences generated as a result of customer interaction with the service provider. Existing churn prediction methods like decision tree typically just classify customers into churners or non-churners while completely ignoring the timing of churn event. Given histories of other customers and the current customer's data, the presented model proposes a new k nearest sequence (kNS) algorithm along with temporal sequence fusion technique to predict the whole remaining customer data sequence path up to the churn event. It is experimentally demonstrated that the new model better exploits time-ordered customer data sequences and surpasses the existing churn prediction methods in terms of performance and offered capabilities.

1 Introduction

Most of telecom companies are in fact customer-centric service providers and offer to its customers variety of subscription services. One of the major issues in such environment is customer churn known as a process by which a company loses a customer to its competitor. Recent estimates suggest that churn rates in the telecom industry could be anything between 25% to 50% [1]. Moreover average acquisition costs standing at around $400 take years to recoup [1] and are estimated to be between 5 to 8 times higher than average retention costs [2]. In such circumstances, it makes every economic sense to have a strategy to retain customers which is only possible if the customer intention to churn is detected early enough. In the presence of large data warehouses packed with terabytes of data, data mining techniques are being adopted to business applications [3] in an attempt to understand, explain and predict customer interaction with the company that leads to churn. Many churn prediction models are available in the market, however churn is only being modelled statically by analysing customer data and running regression or predictive classification models at a particular time instance [4]. On the research arena the focus is shifted towards more complex classification and nonlinear regression techniques like neural networks [5] or support vector machines [6] yet still applied in the same static context to customer data.

E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 207–215, 2006.

In this work the weaknesses of static churn prediction are addressed resulting in a proposition of a new temporal churn prediction system. It uses novel k nearest sequence algorithm that learns from the whole available customer data path and is capable to generate future data sequences along with precisely timed predicted churn events.

The remainder of the paper is organised as follows. Section 2 formulates the problem of churn prediction and explains the sequential data model used for predictions. Next section shows Gamma distribution-based modelling of customer lifetime as a prerequisite tool for kNS algorithm described in Section 4. Section 5 provides comparative churn prediction results from experiments carried out upon real customer data. Finally the concluding remarks are drawn Section 6.

2 Problem Formulation and Data Model

Assuming a constant stream of N customers from whom n_{churn} customers churn while n_{new} customers join the company in a considered time window let $n_{churn} = n_{new}$ and let $p_{prior} = n_{churn}/N$ stand for a prior churn probability. As only churn predictions are actionable and incur some cost, the performance measures evaluating the model should focus on measuring efficiency of churn predictions. Considering predictability limits in the optimal scenario all churn predictions made are correct, whereas in the worst case i.e. by predicting k churners at random, on average kp_{churn} churners will be correctly recognised at the cost of $(1 - kp_{churn})$ wrong churn predictions. The churn prediction results can be presented in a form of confusion matrix as shown in Table 1. The churn recognition rate evaluating efficiency of churn predictions can be expressed by: $p_{churn} = c_{1|1}/(c_{1|1}+c_{1|0})$. In order to truly reflect the quality of model predictions compared to the random predictions it is sensible to use a gain measure which expresses how many times the predictor's churn rate is better than the prior churn rate p_{prior}:

Table 1. Confusion matrix representation: $c_{0|0}$ - true negatives, $c_{1|0}$ - false negatives, $c_{0|1}$ - false positives, $c_{1|1}$ - true positives, where churn is denoted by 1 and non-churn by 0

Actual\Predicted	NonChurn	Churn
NonChurn	$c_{0\|0}$	$c_{1\|0}$
Churn	$c_{0\|1}$	$c_{1\|1}$

$$G = \frac{p_{churn}}{p_{prior}} = \frac{c_{1|1}(c_{1|1} + c_{1|0} + c_{0|1} + c_{0|0})}{(c_{1|1} + c_{1|0})(c_{0|1} + c_{1|1})} \tag{1}$$

Or in order to scale the gain within the achievable limits of $(1, 1/p_{prior})$ a simple normalisation can be used to finally give the relative gain measure defined as follows:

$$G_r = \frac{p_{churn} - p_{prior}}{1 - p_{prior}} = \frac{c_{1|1}c_{0|0} - c_{1|0}c_{0|1}}{(c_{1|1} + c_{1|0})(c_{0|0} + c_{1|0})} \tag{2}$$

Using the above performance measures it is demonstrated that the successful churn prediction strategy depends on comprehensive exploitation of temporal data. These data are generated on the course of customer interaction with the service provider and can be interpreted as a time-directed trajectories in the multidimensional data space generated from customer events. This way not only information coming in the sheer data is used for prediction but also additional information about data sequentiality patterns. The model assumes periodically collected customer data. The duration of each time interval will form the time resolution for customer behaviour analysis and should be selected inline with company routines, data availability and the required time-resolution of generated predictions. Once this is decided, the elementary data unit x_{cti} represents a value of the i^{th} column (feature), collected at time interval t for the customer identified by c. For each customer c the complete life cycle would be defined by the time ordered sequence of customer data that can be stored in a matrix profile as follows:

$$X_c = \begin{bmatrix} x_{c,1,1} & x_{c,1,2} & \dots & x_{c,1,M} \\ \dots & \dots & \dots & \dots \\ x_{c,T,1} & \dots & \dots & x_{c,T,M} \end{bmatrix} \tag{3}$$

where T is the total number of time intervals of a sequence and M stands for the number of features. To increase storage efficiency customer data matrices were piled together into a single table D with the customer identifier and time interval index moved into the two first identifying columns. This tabular data model is more suitable for SQL processing in databases and allows for instant and effortless update of table D. Due to data availability issues, in our models the time resolution has been set to 1 month such that the annual customer data form an ordered sequence of 12 points.

3 Customer Lifetime

Customer lifetime in itself provides unconditional estimation of the churn timing blind to any additional data describing customer interaction with the service provider. Historical lifetime data from different customers gives information about the lifetime distribution which then allows to link the churn event to a random process with certain probability distribution and hence describe churn in probabilistic terms. Denoting by x_i a random series of customer lifetimes extracted from the available data the first goal is to find its distribution. Customer lifetime can be modelled as a random process modelling random event occurrence where the waiting time (i.e. lifetime) before the event (i.e. churn) is relatively long. The three distributions that match this process have been identified: exponential, Gamma and Weibull distributions. All of these distributions have reported applications to lifetime modelling [7]. Yet further analysis in which distributions were fitted to certain ranges of lifetimes e.g. for lifetimes greater than 5 years showed that only Gamma distribution consistently fits well the data despite its changing characteristics. Settling on the Gamma distribution, the probability density function (PDF) of the lifetime x is defined by:

$$y = f(x|a,b) = \frac{x^{a-1}e^{-x/b}}{b^a \Gamma(a)} \qquad \Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt \tag{4}$$

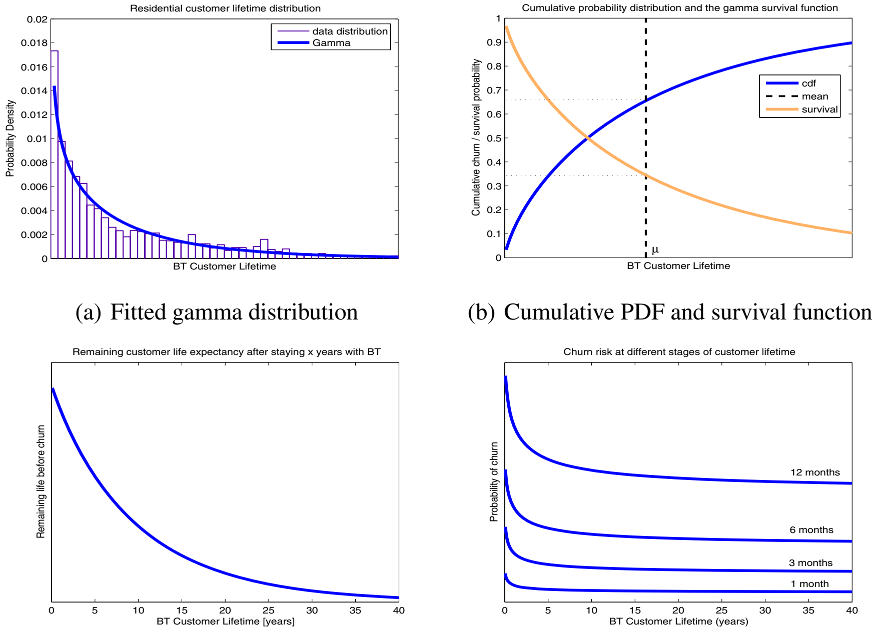
where a and b are the distribution parameters and $\Gamma(a)$ is a gamma function. Given the density (4) the cumulative distribution function (CDF) can be obtained by:

$$p = F(x|a,b) = \int_0^x f(x)dx = \frac{1}{b^a \Gamma(a)} \int_0^x t^{a-1} e^{-t/b} dt \qquad (5)$$

The CDF function $F(x)$ returns the probability that an event drawn from gamma distribution will occur no later than at x. In case of customer lifetime modelling the only event expected is customer churn, hence the CDF becomes immediately the churn risk expressed as a function of the customer life with the service provider. Complementing the event of churn occurrence to a certain event, i.e. $S(x) = 1 - F(x)$ gives the customer survival function showing the probability $S(x)$ of a customer surviving up to the period x. Figures 1(a)-1(b) show examples of the fitted gamma distribution along with corresponding cumulative probability distribution and the survival function.

Exploiting further the properties of the gamma distribution one can immediately obtain the average customer lifetime expectancy and its variability which correspond to the gamma mean $\mu = ab$ and variance $\nu = ab^2$.

The lifetime analysis carried out so far applies to a customer who just joined the service. In reality at any snapshot of a time one might need to ask about the remaining lifetime of a customer who already stayed with the service provider for some time. The problem of a remaining lifetime from the mathematical point of view is quite simple.



(a) Fitted gamma distribution

(b) Cumulative PDF and survival function

(c) Remaining lifetime expectancy curve

(d) Churn risk evolution curves

Fig. 1. Gamma PDF distribution fitted to the customer lifetime data. 1(a) Gamma PDF fitted to customer lifetime data 1(b) Cumulative probability distribution of the fitted gamma distribution. 1(c) Remaining lifetime expectancy curve. 1(d) Churn riskl evolution curves.

The remaining lifetime of a customer at age t can be found by an integration of the survival function from t to ∞ and renormalisation which can be expressed by:

$$L(t) = \frac{1}{S(t)} \int_t^{\infty} S(x)dx \tag{6}$$

Note that if the customer has just joined the company i.e. when $t = 0$, the remaining lifetime simplifies to the Gamma mean parameter: $L(0) = \int_0^{\infty} S(x)dx = \mu$.

Figure 1(c) shows the plot of the remaining lifetime for incremental periods that the customer already stays with the company obtained by numerical integration shown in (6). Finally the most interesting characteristics from the churn prediction point of view is the churn risk evolution over time. Assuming that the churn probability relates to fixed time ahead τ, its evolution over time $H(t)$ can be calculated by the integration of gamma PDF function from t to $t + \tau$ and renormalisation i.e.:

$$H_{\tau}(t) = \frac{\int_t^{t+\tau} f(x)dx}{\int_t^{\infty} f(x)dx} = \frac{F(t+\tau) - F(t)}{1 - F(t)} = \frac{S(t) - S(t+\tau)}{S(t)} \tag{7}$$

Using formula (7) examples of churn risk evolution curves were obtained for 1, 3, 6 and 12 months as shown in Figure 1(d)

The presented analysis indicates that entrants have much higher churn rate than the customers who stayed long with the service provider. For churn prediction purposes this information means that the random churn prediction strategy from previous section would give much higher performance if applied only to entrants or in general to customers with the service age at which the churn risk is the highest.

4 K Nearest Sequence Algorithm

As mentioned in Section 1 temporal classification approach to churn prediction struggles from its underlying static form in which a classifier can only give the answer on whether the churn will occur or not in the particular time slot. Such models either loose information through the necessity of temporal data aggregation or ignore it by not being able to exploit the dynamics of data evolution over time. The timed predictions of churn and its circumstances is beyond the capabilities of temporal classification models.

In an answer to these challenges an original, simple and efficient non-parametric model called k Nearest Sequence (kNS) has been developed. This model uses all available sequences of customer data on input and generates the whole remaining future data sequences up to the churn event as an output. An interesting part is that this model does not require learning prior to the prediction process. Instead it uses customer sequence as a template and automatically finds the k best matching data subsequences of customers who already churned and based on their remaining subsequences kNS predicts the most likely future data sequence up to the time-stamped churn even. The model is further supported by the remaining lifetime estimate presented in Section 3 which is used to add the estimated lifetime feature to the training data.

Let for notation simplicity $S(c, t, \tau)$ stand for a data sequence for a customer identified by c, starting from the time period t and having the length τ i.e.:

$$S(c, t, \tau) = \{\mathbf{x}_{c,t}, \ldots, \mathbf{x}_{c,t+\tau}\} \quad \tau = 1, 2, .. \tag{8}$$

Formally the objective of kNS is to predict the future sequences of customer data $S(c, t_{cur} + 1, R)$, where R is the remaining customer lifetime and t_{cur} is the current time, given the existing customer data sequence to date $S(c, t_{cur} - \tau, \tau)$, where τ is the data availability timeframe, and the historical and live data of former and current customers $S(c_i, 0, L_i)$, where L_i is a lifetime of the customer identified by c_i.

In the first step kNS finds the k nearest sequences, i.e. the customers whose data sequences match the considered sequence the most. This task is split into 2 subtasks. First, all customer sequences are scanned to find the best matching subsequences i.e. the subsequences $S(c_i, t_i - \tau, \tau)$ that have the closest corresponding points to the considered sequence pattern $S(c, t_{cur} - \tau, \tau)$ in the Euclidean distance sense, that is:

$$t_i = \arg\min_{t=\tau}^{L_i} \|S(c, t_{cur} - \tau, \tau) - S(c_i, t - \tau, \tau)\| = \arg\min_{t=\tau}^{L_i} \sum_{j=0}^{\tau} \|\mathbf{x}_{c,t_{cur}-j} - \mathbf{x}_{c_i,t-j}\| \tag{9}$$

Then all best subsequences $S(c_i, t_i - \tau, \tau)$ are sorted according to their distance from the sequence in question in order to determine first k best matching patterns. The remaining subsequences of these best matches i.e. $S(c_i, t_i, L_i - t_i)$, $i = 1, .., k$, are referred to as k nearest remainders, and are used directly for predicting the future sequence in question $S(c, t_{cur}, L)$. The prediction is done by a specific aggregation of the k nearest remainders, which due to different lengths of the remaining subsequences has been supported by the time stretch of the sequences such that they all terminate in the same average churn point. First, the remaining lifetime is calculated by taking average from the last points of k nearest remainders:

$$R = \frac{1}{k} \sum_{i=1}^{k} L_i - t_i \tag{10}$$

Then denoting by $s_i = (L_i - t_i)/R$, $i = 1, .., k$ the transition step for each of k nearest remainders, each j^{th} point $\mathbf{x}_{c,t_{cur}+j}$, $j = 1, .., R$ of the predicted sequence can be calculated using the following formula:

$$\mathbf{x}_{c,t_{cur}+j} = \frac{1}{k} \sum_{i=1}^{k} \left[(\lceil js_i \rceil - js_i) x_{c_i,t_i+\lfloor js_i \rfloor} + (js_i - \lfloor js_i \rfloor) x_{c_i,t_i+\lceil js_i \rceil}\right] \tag{11}$$

which uses interpolation of in-between sequence points spaces to obtain higher precision when dealing with unequally lengthened sequence reminders. This way kNS algorithm would predict the churn taking place in $L = t_{cur} + R$ time period along with the data path $S(c, t_{cur}, R)$ leading to this event.

In realistic scenario the data availability timeframe is very narrow compared to customer lifetime, which means that instead of having complete data sequences of customers from their acquisition up to churn there is only a tiny time slice of data available. In this time slice the turnover of customers and hence the number of churners that can be observed is on average $\tau N/L$ and . Most of the data in such It is therefore expected that vast majority of data relate to the existing customers who did not churn yet. In this case it is very likely that among k nearest sequences most would relate to existing customers with unknown churn time. In this case the kNS algorithm is aided by the unconditional remaining lifetime estimate covered in Section 3. The data sequence predictions are provided only up to the data availability limit and then overall customer average sequence is filled in up to the predicted churn event. The complete kNS algorithm is visualised in Figure 2.

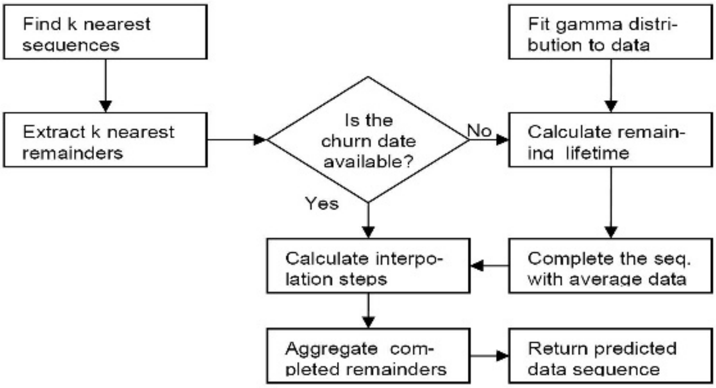

Find k nearest sequences
Extract k nearest remainders
Is the churn date available?
No
Yes
Fit gamma distribution to data
Calculate remaining lifetime
Complete the seq. with average data
Calculate interpolation steps
Aggregate completed remainders
Return predicted data sequence

Fig. 2. Flowchart of the k nearest sequence algorithm

5 Experiments

In order to demonstrate the capabilities of the presented kNS algorithm a set of comparative experiments have been carried out on the real BT customer data sample. Due to data security issues only selected features describing customer provisions, fault repairs and complaints events were used excluding possibly crucial but sensitive customer billing data. They were prepared compliant to data model presented in Section **??** with first column of customer identifier, second keeping monthly time period, third holding customer lifetime up to the current month, and fourth keeping the churn label i.e. a binary flag indicating whether the customer churned in this month or not. This follows with 24 columns of customer data features extracted from customer events database and including events counters, durations and many other mostly quantitative measures describing these events. The data were sampled at random from the database but due to very narrow data availability timeframe of just 10 months churners have been selected for only last month in order to keep customer sequences the longest possible and of consistent length. All the customers data selected for the experiment had at least one event of each type. As a result about 20000 10-months customer sequences have been extracted with many missing data. These missing data have been treated using exponential continuous decay function of the form $y(t) = e^{-ln(2)t/T}$, where T stands for the half-decay period, which intends to simulate fading intensity of human emotions caused by certain event over time. The 2-weeks half-decay period was used following some quick manual optimisation. The prior churn probability i.e. the average likelihood that random customer will churn in particular month turned out to be very low and is denoted by p_{prior} as summarised in Section 2. Due to security reasons the real value of p_{prior} can not be shown here and for the same reasons in all experiments the model performance is expressed by a gain measure as described in Section 2.

The first experiment concerned predicting customer lifetime based on lifetime analysis presented in Section 3. Based on gamma distribution fitted to customer lifetimes, the period with the highest churn density was identified to be the first year of contract and accordingly random churn prediction was applied to the customers who stay less than a year at the time of measurement. The resultant performance brought a gain of just $G_{lifetime} = 1.46$ achieving 46% improvement in correct churner prediction.

In the next experiment a number of standard classifiers were used to classify customer states in the last month (Oct-2004) only using 10-fold cross-validation testing method. The results are shown in the Table 2.

Table 2. Gain measures obtained for the 3 linear classifiers applied to the last month data

Classifier	Dec. Tree	LDA	Fisher	Quadratic	FF NNet	SVM
Gain	11.28	10.24	10.58	10.86	9.97	11.06

Finally the presented kNS algorithm was tested using all 20000 10-months customer sequences with approximated missing data when necessary. Due to the fact that kNS returns precise churn timing it has been assumed that correct churn recognition occurs when the predicted timing deviates no more than 2 months from the actual churn date. The experiment was run for 5 different nearest sequence parameters $k = 1,..,5$ and for 5 different lengths of the matching template i.e. for $\tau = 1,..,5$ months. For each setup the performances have been converted to the gain measures which are shown in Figure 3 along with the diagram depicting differences between these measures. The results show very clearly that above $k \geq 3$ performance gain obtained for kNS becomes higher than for any other tested classifier, shown in Table 2. When $\tau = 1$ the algorithm converges to a standard kNN algorithm for which the performance is comparable to other classifiers from Table 2. The sequential strength starts to take effects from $\tau > 1$ and the results shown in Figure 3 confirm that the highest performance gains are observed for τ ranging between 2 and 4 months with the highest gain of 13.29 obtained for 3-months sequences matched to 4 nearest sequences. These results outperform static non-sequential classification by around 20% and thereby confirm the contribution of the sequential modelling to the churn prediction performance. Due to the lack of space the whole issue of data sequences prediction, coming as a by-product of kNS-based churn prediction, was ignored and left for further investigations within a wider framework of of customer behaviour modelling.

6 Conclusions

Summarising, this work uncovered a new perspective on customer churn prediction and highlighted the importance of the sequential analysis for events prediction. The traditional methods of churn prediction based static classification have been shown conceptually unable to handle the sequential nature of customer relationship path up to churn and therefore unable to time the predictions. The presented kNS algorithm was

$k \setminus \tau$	1	2	3	4	5
1	6.53	6.53	8.05	7.37	7.58
2	8.80	11.20	10.71	7.61	9.06
3	9.00	11.25	10.75	12.87	9.09
4	10.62	12.82	**13.29**	13.18	12.87
5	11.05	13.02	12.35	12.33	12.42

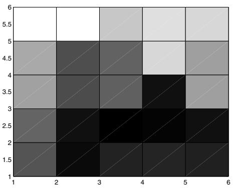


Fig. 3. Gain performance measures obtained for kNS predictor for 5 different nearest sequence parameters k and template sequence lengths τ, shown in a tabular form (left) and graphical diagram (right). Darker shades correspond to higher gain measures.

designed specifically to learn from customer data sequences and is capable to handle the whole customer life cycle rather than individual customer state caught in a snapshot of time. The kNS algorithm is prepared for limited data availability timeframe and can effectively handle missing data. Moreover kNS is capable of exploiting both former customers with completed lifetime data paths and the existing customer sequences by using Gamma distribution applied to model expected customer lifetime. Due to the lack of space and the churn focussed theme of this work the whole aspect of data sequences prediction, coming as a by-product of kNS-based churn prediction, was ignored and left for further investigations within a wider framework of customer behaviour modelling.

References

1. G. Furnas. Framing the wireless market. The Future of Wireless, WSA News:Bytes 17(11):4-6, 2003.
2. L. Yan, D.J. Miller, M.C. Mozer, R. Wolniewicz. Improving prediction of customer behaviour in non-stationary environments. In Proc. of Int. Joint Conf. on Neural Networks, 2001, pp. 2258-2263.
3. M. Morgan. Unearthing the customer: data mining is no longer the preserve of mathematical statisticians. Marketeers can also make a real, practical use of it (Revenue-Generating Networks). Telecommunications International, May, 2003.
4. R.O. Duda, P.E. Hart, D.G. Stork. Pattern classification. John Wiley & Sons, 2001.
5. M. Mozer, R.H. Wolniewicz, D.B. Grimes, E. Johnson, H. Kaushansky: Churn Reduction in the Wireless Industry. Neural Information Processing Systems Conf. 1999, pp 935-941.
6. K. Morik, H. Kopcke: Analysing customer churn in insurance data. Eur. Conf. on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, 2004, pp 325-336.
7. M.S. Finkelstein. On the shape of the mean residual lifetime function. Applied Stochastic Models in Business and Industry 18(2):135 - 146, 2002.