Predictive Power of Online and Offline Behavior Sequences: Evidence from a Micro-finance Context

Rishabh Mehrotra  
*University College London, erishabh@gmail.com*

Prasanta Bhattacharya  
*Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), news.aaruush@gmail.com*

Tianhui Tan  
*national university of singapore, tinweitam@gmail.com*

Tuan Phan  
*National University of Singapore, tuan.quang.phan@gmail.com*

Follow this and additional works at: [http://aisel.aisnet.org/icis2017](http://aisel.aisnet.org/icis2017)
Predictive Power of Online and Offline Behavior Sequences: Evidence from a Microfinance Context

Short Paper

Rishabh Mehrotra
Department of Computer Science,
University College London
London, United Kingdom
r.mehrotra@cs.ucl.ac.uk

Prasanta Bhattacharya
Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore,
1 Fusionopolis Way, Singapore 138632
prasanta_bhattacharya@ihpc.a-star.edu.sg

Tianhui Tan
Department of Information Systems and Analytics,
National University of Singapore
15 Computing Drive, Singapore, 117418
distt@nus.edu.sg

Tuan Q. Phan
Department of Information Systems and Analytics,
National University of Singapore
15 Computing Drive, Singapore, 117418
disptq@nus.edu.sg

Abstract

Microfinance based institutions have emerged as a potential solution to the financial exclusion problem in developing economies around the world. A key challenge facing such micro-lending firms is assessing the credit risk of borrowers, owing to the lack of formal financial histories and collaterals. A number of micro-lending companies have, therefore, started leveraging social media and digital communication data from applicants to assess their ability and willingness to repay loans. In our study, we demonstrate a novel approach of leveraging online and offline behavior sequences, as captured from the borrowers’ browsing logs and mobility traces to accurately predict the borrowers’ creditworthiness. Our preliminary results show that using such sequence data, we can provide micro-lending firms with a cheap and reliable strategy for assessing credit risk of borrowers at the time of loan creation. We contend that such big-data based strategies are critical to the sustainability of micro-lending institutions.

Keywords: Microfinance, location, browsing, LSTM

Introduction

Financial exclusion has been widely recognized as a major hindrance to socio-economic development in developing countries worldwide (Demirgüç-Kunt et al. 2015), and refers to a situation where the poor and disadvantaged are prevented from accessing financial services, like seeking credit. Recent reports on the
Predictive Power of Behavior Sequences

topic suggest that as many as 38% of the world’s population still do not have access to formal banking institutions. As a result, a number of developing countries have organically developed semi-formal and informal institutions for microfinance (Yunus 2007) and peer-to-peer lending (Hartley 2010). However, there hasn’t been much work on studying the efficacy and risks facing these institutions partly because of the unavailability of high-quality data on the borrowers’ financial history and demographics. However, the high penetration of mobile devices and internet access (GSMA 2015) in these countries offer a new and unparalleled source of fine-grained user behavior data. As a result, some organizations in these countries have been leveraging emerging data-driven technologies, like smartphones, to offer banking and credit infrastructures to the resource poor (Dwoskin 2015; Green 2014; Groenfeldt 2015). In our current study, we leverage recent advances in deep learning based sequence models and also fine-grained data on loan borrowers’ online and offline behavioral sequences (e.g. online browsing and offline traveling) as obtained from their mobile phones to predict their credit-worthiness. We specifically highlight that the importance of sequential data on human behaviors from online and offline channels as being highly predictive of credit-worthiness in such developing economies, where other sources of data might be harder to obtain.

Previous studies have hinted at the importance of considering borrowers’ physical mobility patterns in predicting various behavioral outcomes such as traveling styles (Cho et al. 2011), subsequent venues (Noulas et al. 2012), choice of making friends (Scellato et al. 2011b) and credit scores in developing communities (Tan et al. 2016). While Tan et al. (2016) focus on the frequency of visits to specific locational categories as good locational features, we believe it is also important to focus on the specific sequence of location visits, as these sequences might be indicating of finer heterogeneities among the borrowers. For example, one can easily imagine that a person who visits a “school” and a “health” clinic on alternate days for a month, exhibits a very different behavioral profile from a person who visits the school for the first 15 days and the health clinic for the final 15 days of the month. However, a frequency-based approach to modeling mobility over time would fail to capture these behavioral differences which might have associations with the borrowers’ ability of willingness to repay the loans (e.g. in the second case of the example, the borrower might have been hospitalized for the last 15 days, and might hence default on her loan owing to high medical costs). In our study, we adapt a set of state-of-art deep learning architectures, namely Long-term-short-term memories (LSTMs) that are specially optimized for modeling sequence data, to fit two sets of sequences about borrowers’ online behavior as captured by their daily website visits as well as their offline behavior as captured by their physical locational visits. We apply this model to data obtained in collaboration with a pro-social lending app from a major South East Asian country that is a fast-growing economy with very high mobile phone and internet penetration. Preliminary results from our analysis show that such behavioral sequences alone offer a good predictive accuracy in determining creditworthiness of individuals, and offer superior performance when compared to other baseline methods that use locational information but ignore sequential nature of it, as with Tan et al. (2016).

This work-in-progress paper is among the first to highlight the role of understanding and analyzing different types of behavioral sequences that are easily obtainable from mobile phone records, in predicting creditworthiness in those emerging economies which lack formal banking infrastructures. We emphasize how recent advances in sequence-based deep learning models can be intelligently leveraged to fit both online as well as offline sequences of human behavior. Further, we show that browsing behavior and mobility behavior exhibited by the borrower in the post-loan period is a strong predictor of a user’s willingness to pay back a loan. Such insights are critical to microfinance and lending institutions, as well as to governments and policy makers interested in implementing financial services for the resource poor.

In the following section, we present a brief literature review about previous work on sequence-related data, focusing specifically on browsing-related and mobility-related behaviors. Following this, we introduce the architecture of our LSTM models that we have designed for our specific context. Next we introduce our empirical and data context, and present some preliminary results from our analyses. Finally, we highlight the insights we generate from our current analyses, and discuss potential future directions for this study.
Background

Sequence

The importance of sequence has long been recognized in various domains including natural language processing (Manning and Schütze 1999) and computational biology (Durbin et al. 1998). Moreover, by leveraging temporal sequence data, we can also discover and predict user-level behavioral patterns for various contexts. For instance, with a belief that people tend to repeat themselves consciously or unconsciously, Davison and Hirsh (1998) conceptualized the user behavior prediction task as predicting the next element in a sequence made up of nominal elements. They proposed a generic learning algorithm which required little domain knowledge, and demonstrated its efficient performance in command input prediction. In other work, Pentland and Liu (1999) regarded humans as a device of many internal states, and drew on this insight to develop dynamic Markov models to predict subsequent behaviors based on initial behaviors. In order to capture the changes in latent states and uncover the behavior chain, Hidden Markov models are now often applied in a multitude of applications (Rabiner 1989). In recent years, more advanced methods have been proposed to predict user complex behaviors based on temporal sequence patterns, specifically in the context of medical record histories, web click logs and movie viewing (e.g. Koren 2010; Li and Fu 2014; Matsubara et al. 2012; Yang et al. 2014). As a result of the superior prediction performance, sequence based models are widely applied in real-world practices such as targeting advisements and recommendation system. However, most of these studies focus on sequential behaviors of the same kind. There is limited evidence on whether these sequential behaviors of one kind could act as a predictor for a behavior of a different kind. In other words, little is known about whether the sequence of online browsing behaviors or the sequence of offline visiting behaviors could be correlated to financial behavior, i.e. cash withdrawals or loan repayments.

Mobility

With the rapid penetration of information and communication technologies (ICT), like GPS-enabled smartphones, a user’s mobility now can be easily captured. Hence, there is an increasing interest in using spatio-temporal features to predict user-level behavioral outcomes, due to the massive amount of fine-grained sensor data. Most of the studies make use of locational information (i.e. “types”) from check-ins on apps like Foursquare and Gowalla (Cho et al. 2011; Noulas et al. 2011). The behavioral outcome could be the traveling styles (Cho et al. 2011), bandwidth-based QoS (Evensen et al. 2011) as well as user’s next location (e.g. Monreale et al. 2009; Noulas et al. 2012). For example, assuming that the sequence of daily visited locations is consistent, Scellato et al. (2011a) proposed a spatio-temporal framework to predict a user’s next place to visit using his/her arrival time and residence time. With the spatio-temporal data and models available, business owners can now harness their consumers’ mobility for business objectives such as targeting and evaluation.

Browsing

Web browsing is another behavior that attracts large attention since ICT, including personal computers and smartphones, have become pervasive. For example, leveraging data on clickstream, search logs, etc., researchers have shown that navigational patterns can be categorized and linked to purchase likelihood (Moe 2003) and also used to predict future movements at a web site (Montgomery et al. 2004). Other user-level behavioral outcomes include the choice to stay within the site (Bucklin and Sismeiro 2003), response to banner advertisements (Chatterjee et al. 2003) or even decision to MOOC dropout (Kloft et al. 2014). Business owners could have access to the huge amount of clickstream data to understand the browsing patterns of their consumers, which helps to better realize personalized marketing.

As mentioned earlier, in our study we leverage sequences of online browsing and offline location visits to help predict loan repayment behavior. The following section presents the design of the location and browsing based sequence model that we developed.
Predictive Power of Behavior Sequences

Modeling Sequence Information using LSTMs

Location Sequence Model

Recurrent neural networks (RNNs) are a family of neural networks that encodes informational dependencies learned over past events, and uses such information to reason about current ones.

RNNs achieve such a connectionist properties using loops or cycles in the network which allows past information to persist in the form of inputs to the next network. In a way, RNNs can be thought of as a sequence of the same neural network where each unit passes the message on to the next unit after applying set of weights and transformations. This chain-like characteristic of RNNs makes them naturally suitable for data in the form of sequence or lists. In recent years, RNNs have been used successfully in a variety of contexts (Graves 2012; Graves et al. 2006; Lipton et al. 2015).

While, in theory, RNNs are considered good at encoding long-term dependencies in their networks, they often tend to perform poorly owing to the vanishing gradient problem or the exploding gradient problem (Bengio et al. 1994; Pascanu et al. 2013). The vanishing problem appears frequently with networks with gradient based methods wherein, and owing to the choice of specific activation functions, the gradient becomes smaller as one moves back to the earlier layers. This means that neurons in the earlier layers learn much more slowly than neurons in later layers. The opposite problem, where earlier layers learn faster, is called the exploding gradient problem. LSTM networks (Hochreiter and Schmidhuber 1997) are an improvement over RNNs in the sense that they can better cope with gradient instability problems. In its simplest form, a LSTM unit comprises three multiplicative gates which decide what proportion of information passes on to the next time step, and consequently, what proportion of information gets lost or forgotten. Such LSTM units together with other LSTM and non-LSTM units form an entire LSTM network. It serves as a recurrent model that remembers or stores outputs for shorter or longer durations. It avoids the gradient instability problem since these networks use no activation function within the recurrent components and hence the gradient does not change significantly during backpropagation over time. Instead, LSTM units are generally implemented in “blocks” having multiple LSTM units. Formally, the formulas to update an LSTM unit at time t are as follows:

\[
\begin{align*}
i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\
f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\
\tilde{c}_t &= \tanh(W_c h_{t-1} + U_c x_t + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\
h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]

where, \(\sigma\) is a sigmoid function, \(\odot\) is the element-wise product operator, \(x_t\) is the input vector at time \(t\), \(h_t\) is the hidden state or output vector storing all useful information up until time \(t\). The weights of different gates for input \(x_t\) are denoted by \(U_i, U_f, U_c, U_o\), and the weights for the hidden state \(h_t\) are denoted by \(W_i, W_f, W_c, W_o\). Also, \(b_i, b_f, b_c\) and \(b_o\) refer to the bias vectors. Given the nature of the locations users visit over time, we model such sequential location information as an LSTM. Figure 1 below illustrates the neural network architecture used to encode the sequential location information.
URL Sequence Model

We leverage hierarchical sequence models to encode URL sequences visited by a given user over time. We introduce an LSTM model that hierarchically builds an embedding for a URL sequence from embeddings for URLs from their title text. Unlike location sequences, URL sequences are nested in the sense that the sequence of URLs is composed of URL titles, and each URL title has a sequence of words. To encode this hierarchically nested url-title information, we leverage a hierarchical LSTM model. For each URL, we make use of its Title information, and learn an LSTM model over word sequences in the url-title.

As shown in Figure 2 below, we design a hierarchical LSTM for URL-sequence modelling with two layers: the first layer receives title words as inputs and generates word-level representations (Title-LSTM) for each URL, and the second layer takes the URL-level representations as inputs and yields user-level representations (URL-LSTM) obtained from URL sequence. By means of the provided end-of-word labels, the Title-LSTM obtains word-level representation after processing the last word of each title and passes the word-level representation to the URL-LSTM. Thereafter, the URL-LSTM performs an update of the URL-level representation. Note that the hidden states of the URL-LSTM remains unchanged while all the words of a title are processed by the Title-LSTM. When the URL-LSTM starts to process the next URL, its hidden states are reinitialized using the latest hidden states of the URL-LSTM, which contain summarized representation of all the URLs that have been processed by that time step, in that phrase. The output of the URL-LSTM is then fed into the final classifier layer of the model which is used for differentiating loan defaulters.

The hierarchical URL-LSTM architecture was implemented in Keras with Tensorflow backend on a GPU cluster. The model was jointly trained in an end-to-end learning fashion wherein the embedding layer as well as the recurrent layer were jointly trained based on the final misclassification loss.
Data Description

We collaborate with a microfinance company which offers micro loans in a major Southeast Asian
country. The country is categorized as a lower middle income economy although it has one of the world’s
largest populations. It does not have any formal credit scoring bureaus or credit reporting systems and
hence, people find it difficult to obtain access to credit without collaterals, which has led to growing
financial exclusion in the country. On the other hand, the country has experienced a fast growing Internet
population and mobile phone penetration in recent years. As reported, more than 40% of the population
now has access to a smartphone. Leveraging this opportunity, the microfinance company has started an
online and app-based loan issuance business, with minimal requirements on offline paperwork. It
launched an Android app in 2014, to allow borrowers to submit loan applications via their mobile devices.
To facilitate the loan interest generation, the app collected phone related data such as GPS locations,
voluntarily shared by the borrowers. Thus, other than possible collaterals and/or their financial history,
borrowers are actively encouraged to share their phone details and/or SNS (e.g. Facebook) accounts with
the company to obtain better credit scores, generated from these data points using a proprietary
algorithm. On average, the company has offered small loans equivalent of USD 300 for a duration of 3-6
months to the emerging middle class in the country.

We obtained anonymized backend data for a sample of users who had borrowed a loan in 2014. Our data
set consists of two components: (i) loan details such as loan amount, interest rate and the
repayment/default information, which is our focal outcome variable, and (ii), online and offline behaviors
of the borrowers recorded from their mobile phones, which includes phone usage data such as installed
apps and saved contacts, communication data such as phone calls and text messages, web browsing data
such as the sequential URLs visited, and geo-location data such as GPS mobility traces.

For the current study, we focus on whether and how sequential behaviors like web-browsing and mobility
patterns can be intelligently leveraged to predict financial outcomes, like loan repayment. This is
consistent with recent work that has emphasized how mobile phone information can prove to be an
effective data source for loan repayment prediction (e.g. Bjorkegren and Grissen 2015; Tan et al. 2016), a
view that is also consistent with the pragmatism idea of credit scoring (Thomas et al. 2005).

The mobility data is in the form of geocodes that are coded as pairs of longitudes and latitudes. In order to
extract meaningful abstractions from this data, we used the Google Maps API to reverse map the original
geocodes to specific location/property types, from within a certain radius. The Google Maps API supports a comprehensive list of location types including neighborhood, hospital, airport, café etc. To compute the sequence chains of location visits for each user, we focused our attention to all location visits by the user for a period of 7 days immediately following the loan-issuance process. We concatenated these location categories to construct a chain of location types (e.g. “school->hospital->airport->home->home->airport”) for each borrower in our dataset. A similar sequence was constructed using URL and topic data that we extracted from the borrowers’ browsing logs from the app. Similar to the strategy mentioned in the previous paragraph, we concatenated all websites and website categories for a period of 7 days immediately following the loan issuance to construct a sequence of URLs and a sequence of URL topics for each user. The basic intuition behind using sequence data for a period of 7 days following the loan issuance process is to exploit immediate changes in behavior resulting from the economic shock (i.e. obtaining an extra sum of money from the loan), to make accurate predictions of credit-worthiness. It is important to note here that the mobility behavior of the users prior to loan issuance maybe different from that post-issuance. However, owing to data limitations, we can only track users in the post-loan period. We do contend, however, that mobility patterns might be reflective of the borrowers’ inner dispositions and habits. In that respect, these patterns might show variance among users (e.g. a heavy spender might visit restaurants and casinos, but not a cautious spender), but not within borrowers over time. Thus, our mobility patterns might be capturing not just the effects of receiving the loan amount, but also individual level differences which tend to be quite stable over time.

<table>
<thead>
<tr>
<th>Time</th>
<th>Logit Model</th>
<th>Locational LSTM</th>
<th>URL LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Days</td>
<td>66.71%</td>
<td>79.55%</td>
<td>76.19%</td>
</tr>
<tr>
<td>14 Days</td>
<td>65.23%</td>
<td>78.12%</td>
<td>75.89%</td>
</tr>
<tr>
<td>21 Days</td>
<td>62.89%</td>
<td>71.11%</td>
<td>75.72%</td>
</tr>
</tbody>
</table>

*based on a 20% hold out test

<table>
<thead>
<tr>
<th>Time</th>
<th>Locational LSTM</th>
<th>Randomized Locational LSTM</th>
<th>Bidirectional Locational LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Days</td>
<td>79.55%</td>
<td>63.89%</td>
<td>65.91%</td>
</tr>
<tr>
<td>14 Days</td>
<td>78.12%</td>
<td>61.75%</td>
<td>62.5%</td>
</tr>
<tr>
<td>21 Days</td>
<td>71.11%</td>
<td>63.24%</td>
<td>68.89%</td>
</tr>
</tbody>
</table>

*based on a 20% hold out test

### Preliminary Results

We highlight some preliminary results in Tables 1 and 2 from the implementation of the proposed Title-LSTM and URL-LSTM neural network architectures, along with few baselines from competing approaches. We report the prediction accuracy in each case and run the experiment over 3 different time frames, incorporating historic user location and browsing data spanned over 7 days, 14 days and 21 days respectively. Based on the browser and location sequence data, the average sequence length of 139.6 locations was observed with a maximum of 268 locations and median of 14 locations. For the URL sequence information, we observed a maximum of 340 URLs with a mean and median of 135 and 132 URLs respectively. As shown in Table 1, we observe that the proposed sequential model outperforms the logit model which does not take sequential data into consideration by over 10% in terms of predictive accuracy. The Location LSTM performs slightly better than the URL LSTM which hints that location sequences are more informative than browsing sequences.

Table 2 compares the performance of the proposed sequential models against a randomized model wherein we randomize the original sequence information and feed it into the location sequence LSTM. We observe a major dip in the performance accuracy which highlights the importance of considering sequence
structure while predicting credit worthiness. Additionally, we compare against Bidirectional LSTMs, a variant of LSTM model which considers the input sequence in both directions. The dip in performance for Bidirectional LSTMs further support our conjecture that it is the forward sequence information that is more helpful in predicting credit worthiness as compared to randomized or reversed sequential information.

Future Work

As next steps in this research, we hope to make significant advances over what has been described in this paper. We intend to jointly encode auxiliary information about the users and their social context with the sequential model and propose a unified coupled model which leverages the sequential structure of the data alongside traditional features known to be predictive of credit worthiness. Additionally, we wish to investigate users' browsing information in detail and extract meaningful abstractions of their digital footprints in terms of the kind of tasks users perform and the breadth of topics they're most interested in. Thirdly, we seek to better understand the theoretical rationale behind the association between the different location types and browsing history and the loan-default behavior. Lastly, we will attempt to benchmark the accuracy metrics from our approach against other popular methods for credit-scoring that leverage similar types of borrower data. However, we would like to point out that the objective of our study is not to propose a credit scoring technique that out-performs existing methods. Instead, we are proposing a method that offers reasonably good accuracy but in data-scarce environments, where conventional data like demographics and financial history might not be available.

One limitation of the current work pertains to the generalizability of our proposed approach to the bigger microfinance population. Since the population of microfinance users would also include those who are more privacy conscious and might not want to trade SNS and phone data in return for a better credit score, it is plausible that the insights from this study might not directly apply to these users. However, that would be a practical concern caused by insufficient data, and not a methodical one.

Concluding Remarks

Behavior sequences are increasingly common in online as well as offline contexts, and prior studies have leveraged temporal dimensions, in the form of temporal sequences of information, to make accurate predictions in various contexts. In this study, we contend that such sequence of behaviors could hold predictive value in microfinance contexts as well, where the willingness to repay loans is highly subjective and cannot be formally assessed at the time of loan issuance. We draw on state of art advances in deep learning based LSTM models to predict the loan default behavior of a sample of borrowers based on their online browsing sequence as well as offline locational sequences. Our preliminary results confirm our assertion that such behavioral sequences hold high predictive value in creditworthiness assessment. Specifically, we find that locational sequences outperform other comparable models by a significant margin in predicting the default behavior. The findings from our study contributes to a better understanding of behavior chains in the financial contexts, while also helping micro-lending companies better assess credit risks at the time of loan generation.

References


