SBP-BRiMS 2017

Mitigating the risks of financial exclusion: Predicting illiteracy with standard mobile phone logs

Pål Sundsøy

Telenor Group Research, Big Data Analytics
Snarøyveien 30,1331 Fornebu, Norway

Abstract. The present study provides the first evidence that illiteracy can be predicted from standard mobile phone logs. By deriving a broad set of novel mobile phone indicators reflecting users' financial, social and mobility patterns this study addresses how supervised machine learning can be used to predict individual illiteracy in an Asian developing country, externally validated against a large-scale survey. On average the model performs 10 times better than random guessing with a 70% accuracy. Further it reveals how individual illiteracy can be aggregated and mapped geographically at cell tower resolution. In underdeveloped countries such mappings are often based on out-dated household surveys with low spatial and temporal resolution. One in five people worldwide struggle with illiteracy, and it is estimated that illiteracy costs the global economy more than $1 trillion dollars each year. These results potentially enable cost-effective, questionnaire-free investigation of illiteracy-related questions on an unprecedented scale.

1 Introduction

Illiterates are often trapped in a cycle of poverty with limited opportunities for income generation and financial inclusion. Literacy is also often a hurdle to bring financial services to the unbanked [*1*]. High-quality literacy statistics is therefore crucial to pinpoint areas where better education is needed: where are the illiterates? Mapping of literacy statistics is currently based on tedious household surveys with a low spatial and temporal frequency [*2*]. The increasing availability and reliability of new data sources, and the growing demand of comprehensive, up-to-date international literacy data are therefore of high priority. One of the most promising rich Big Data sources are mobile phone logs (CDRs) [*3*]. CDRs have shown to provide useful proxy indicators for assessing regional poverty levels [*4*,*5*], socioeconomic status [*6*], unemployment [*7*,*8*], infectious diseases [*9*] and disasters [*10*]. This study demonstrates how individual and regional illiteracy can be mapped using a combination of CDRs and financial airtime transactions.

The rest of this paper is organized as follows: Section 2 describes the methodological approach, including the features and modelling approach, while section 3 addresses the research results, followed by concluding remarks in Chapter 4.

2 Approach

2.1 Data

Household survey data: : Data from two nationally representative cross-sectional household surveys of 200,000 individuals in a low-income South Asian country is analyzed. The data is collected at time Q114 and Q214 by an external survey company commissioned by the operator. The survey discriminates between 6 types of educations for the head of household, including being illiterate. The sample includes 6.8% illiterates, 40% primary degree, 26% SSC, 17.6% HSC, 5.6% bachelor, 3.5% master and 0.13% other degrees (incl. Ph.D.). The head of household's is asked for his or her most frequently used phone number. 87% of households in the country has at least one mobile phone.

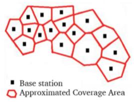
Mobile phone data: Mobile phone logs for 76 000 of the surveyed 200 000 individuals belonging to the leading operator are retrieved from a period of six months and de-identified by the operator. Individual level features are built from the raw mobile phone data and is subsequently coupled with the corresponding de-identified phone numbers from the survey. The social features are subsetted from a graph consisting of in total 113 million subscribers and 2.7 billion social ties. No content of messages or calls are accessible and all individual level data remains with the operator.

2.1 Features

A structured dataset consisting of 160 novel mobile phone features is built, and categorized into three dimensions: (1) financial (2) mobility and (3) social features, as shown in Table 1. The features are custom made to predict illiteracy, and include various parameters of the corresponding distributions such as weekly or monthly median, mean and variance.

Table 1. Sample of features from mobile phone metadata used in model

Dimension	Features
Financial	**Airtime purchases:** Recharge amount per transaction, Spending speed, fraction of lowest/highest recharge amount, coefficient of variation recharge amount etc **Revenue:** Charge of outgoing/incoming SMS, MMS, voice, video, value added sevices, roaming, internet etc. **Handset:** Manufacturer,brand, camera enabled, smart/feature/basic phone etc

Mobility 	Home district/tower, radius of gyration, entropy of places, number of places visited etc.
Social 	**Social Network:** Interaction per contact, degree, entropy of contacts etc. **General phone usage:** Out/In voice duration, SMS count, Internet volume/count, MMS count, video count/duration, value added services duration/count etc.

2.2 Model algorithm

Based on performance of many algorithms, including neural network and SVM, a gradient boosted machines model (GBM) is proposed as the final model [*11*]. To compensate class imbalance, the minority class in the *training set*, containing illiterates, is up-sampled from 6.8%. The minority class is then randomly sampled, with replacement, to be the same size as the majority class. A 10-fold cross-validation is used as re-sampling technique. In this set-up, each model is trained and tested using a 75/25 split. All results are reported for the test-set.

3 Results

3.1 Individual illiteracy

Figure 1 shows the final features and their contribution in predicting illiteracy. Concretely, 19 of the features are related to illiteracy and included in the final GBM classifier. The model predicts whether phone users are illiterate with an accuracy of 70.1% (95% CI: 69.6-70.8). The deviation of accuracy from the training set is only 3.8%, which disregard model overfitting. The true positive rate (sensitivity/recall) is 71.6% and true negative rate (specificity) 70%.

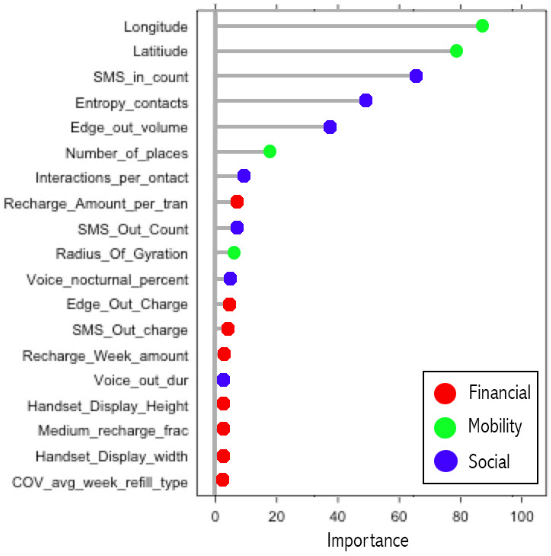


Fig. 1. Top features in the GBM Model colored by their respective feature family

An investigation of the most important predictors, as seen in Fig 1, reveals some interesting associations. We especially notice that most frequently used longitude and latitude stand out as good predictors: where the people spend most of their time is a good signal of their education level. Another important feature is the number of incoming SMS, which (surprisingly) outperforms outgoing SMS. Moreover, we see that entropy of contacts is important – illiterates tend to concentrate their communication on few people. This is also in line with Eagle's work on geographical level [*12*], which shows that economic well-being is correlated with social diversity. Further we see that illiterates have limited use of internet (predictor 5), and their mobility pattern is limited to a few base stations (predictor 6).

3.2 Geographical illiteracy mapping

A natural next step is to move from individual illiteracy to geographical illiteracy. In big Asian cities there are often thousands of mobile towers that can be used as "sensors" to estimate illiteracy rates in the areas covered by the towers. In the rural areas where towers are less dense, interpolation techniques can be utilized to include information from the neighbour towers. Fig 2a) shows the predicted illiteracy rate per tower, in one of the larger cities.

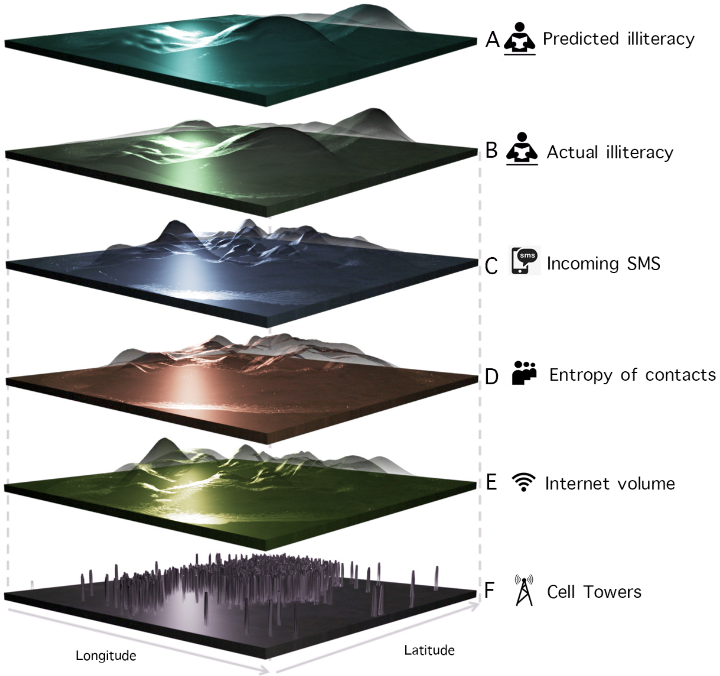


Fig. 2. Geographical mapping of illiteracy, top predictors and the cell tower distribution in one major Asian city. Height (z-axis) is proportional to the tower averages for each given metric.

The individual illiteracy rates are here calculated by using the test set, aggregated and averaged to tower level, and then further spatially interpolated, using an IDW algorithm, to average out the noise of local variations between towers. The *actual* illiterate rates in Fig 2b is calculated by using the training set as ground truth. We notice three large pockets of larger illiteracy rates in the city. By also including distributions of the top predictors (Fig 2 c-e) it is possible to visually observe spatial correlations. For example, one can observe a large area of high SMS activity (Fig 2c,left) that can be associated with low illiteracy rates.

4 Conclusion

This study shows how illiteracy can be predicted from mobile phone logs, purely by investigating users' metadata. By deriving economic, social and mobility features for each mobile user we predict individual illiteracy status with 70% accuracy. Further we show how individual illiteracy can be aggregated and mapped geographically with high spatial resolution on cell tower level. Feature investigation indicates that home cell tower and incoming SMS are the superior predictors, followed by diversity of communication partners and Internet volume. An important policy application of this work is the prediction of regional and individual illiteracy rates in underdeveloped countries where official statistics is limited or non-existing.

References

1. Chibba, M.: Financial inclusion, poverty reduction and the millennium development goals. The European Journal of Development Research 21(2), 213-230 (2009)
2. IHSN: How (well) is Education Measured in Household Surveys? IHSN working paper 002 (2009)
3. Lokanathan, S., Lucas Gunaratne, R.: Behavioral insights for development from Mobile Network Big Data: enlightening policy makers on the State of the Art. Available at SSRN 2522814. (2014)
4. Blumenstock, J., Cadamuro, G. ., On, R.: Predicting poverty and wealth from mobile phone metadata. Science 350(6264), 1073-1076 (2015)
5. Steele, J. E., Sundsøy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y. A., Iqbal, A., Hadiuzzaman, K., Lu, X., Wetter, E., Tatem, A., Bengtsson, L.: Mapping Poverty using mobile phone and satellite data. Journal of The Royal Society Interface 14(127), 20160690 (2017)
6. Sundsøy, P., Bjelland, J., Reme, B. A., Iqbal, A., Jahani, E.: Deep learning applied to mobile phone data for Individual income classification. In : ICAITA (2016)
7. Toole, J. L., Lin, Y. R., Muehlegger, E., Shoag, D., González, M. C., Lazer, D.: Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface 12(107), 20150185 (2015)
8. Sundsøy, P., Bjelland, J., Reme, B. A., Jahani, E., Wetter, E., Bengtsson, L.: Estimating individual employment status using mobile phone network data. arXiv preprint :1612.03870 (2016)
9. Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., Buckee, C. O.: Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences 112(38), 11887-11892 (2015)
10. Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Canright, G. S., Engø-Monsen, K., Bengtsson, L.: Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication, mobility and consumption patterns during cyclone Mahasen. Climatic Change 138(3-4), 505-519 (2016)
11. Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine. Annuals of statistics, 1189-1232 (2001)
12. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science 328(5981), 1029-1031 (2010)