Prediction of Socioeconomic Levels using Cell Phone Records

V. Soto, V. Frias-Martinez, J. Virseda and E. Frias-Martinez

Telefonica Research, Madrid, Spain
{vsoto,vanessa,jvirseda,efm}@tid.es

Abstract. The socioeconomic status of a population or an individual provides an understanding of its access to housing, education, health or basic services like water and electricity. In itself, it is also an indirect indicator of the purchasing power and as such a key element when personalizing the interaction with a customer, especially for marketing campaigns or offers of new products. In this paper we study if the information derived from the aggregated use of cell phone records can be used to identify the socioeconomic levels of a population. We present predictive models constructed with SVMs and Random Forests that use the aggregated behavioral variables of the communication antennas to predict socioeconomic levels. Our results show correct prediction rates of over 80% for an urban population of around 500,000 citizens.

1 Introduction

The Socioeconomic Level (SEL) is an indicator used in the social sciences to characterize an individual or a household economic and social status relative to the rest of the society. It is typically defined as a combination of income related variables, such as salary, wealth and/or education. As such, the socioeconomic status of an individual or a household is also an indication of the purchasing power and the tendency to acquire new goods. The information provided by this variable is very relevant from a commercial perspective, as adapting the interaction between a company and a potential client considering the purchasing power of the client is a key element for the success of the interaction. Also, from a public policy perspective, socioeconomic levels are typically used to implement and evaluate social policies and study their evolution over time. The relevance of the SEL as a factor to explain a variety of human behaviors and social conditions can be widely found in the literature. These studies present the effects that different socioeconomic levels might have in various scenarios like access to health services [1] or public transportation [2].

National statistical institutes provide socioeconomic information, for particular geographical areas, typically stratified into three levels: high socioeconomic level, middle socioeconomic level and low socioeconomic level. Nevertheless, computing these indicators has some limitations: (1) acquiring the data set of socioeconomic levels for a whole country can be extremely expensive; (2) the census and/or the personal interviews needed to calculate SELs are usually done every 5

to 10 years, thus not being able to capture changes in SEL in a timely fashion and (3) although the socioeconomic data for developed economies is reliable, such information in developing economies is not as available and/or reliable because economic activities usually happen in an informal way. As a result, although SELs are key elements for public policy, computing them remains a costly and time consuming procedure.

Due to its ubiquity, cell phones are arising as one of the main sensors of human behavior, and as such, they capture a variety of information regarding mobility, social networks and calling patterns, that might be correlated to socioeconomic levels. In the literature, we can find reports highlighting these relations. For example, [4] and [5] use cell phone records to study the impact of socioeconomic levels in human mobility, concluding that higher socioeconomic levels tend to have a higher degree of mobility. Similarly, authors in [6] study the relation between socioeconomic levels and social network diversity, and indicate that social network diversity seems to be a very strong indicator of the development of large online social communities.

In this paper we evaluate the use of aggregated cell phone data to model and predict the different socioeconomic levels of a population. These socioeconomic prediction models have two potential applications: (1) from a commercial perspective, they can be used to tailor offers and new products to the purchasing power of an individual and (2) from a public policy perspective, they can be used as a complement to traditional techniques for estimating the socioeconomic levels of a population in order to implement public policies and study their impact over time. The application of predictive socioeconomic models solves some of the limitations that traditional techniques to obtain SELs have: they are not based on personal interviews and thus constitute a cost-effective solution.

2 Preliminaries

In order to create models that are able to predict the socioeconomic levels of the population within a geographical area, we propose to use supervised machine learning techniques applied over cell phone records obtained from cell phone networks. First, we give a brief overview about how these networks work.

Cell phone networks are built using a set of base transceiver stations (BTS) that are in charge of communicating cell phone devices with the network. Each BTS tower has a geographical location typically expressed by its latitude and longitude. The area covered by a BTS tower is called a cell. At any given moment, one or more BTSs can give coverage to a cell phone. Whenever an individual makes a phone call, the call is routed through a BTS in the area of coverage. The BTS is assigned depending on the network traffic and on the geographic position of the individual. The geographical area covered by a BTS ranges from less than 1 km^2 in dense urban areas to more than 3 km^2 in rural areas. For simplicity, we assume that the cell of each BTS tower can be approximated with a 2-dimensional non-overlapping region computed using Voronoi tessellation. Figure 1(left) shows a set of BTSs with the original coverage of each cell, and Figure

1(right) presents its approximated coverage computed using Voronoi. Our final aim is to predict the socioeconomic level of each cell in the Voronoi tessellation using the aggregated cell phone information of the BTS tower that gives coverage to each area.

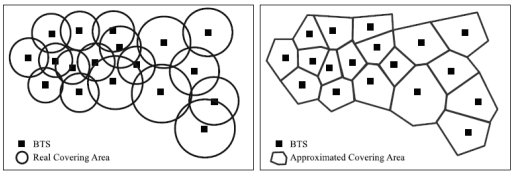

Fig. 1. (Left) Example of a set of BTSs and their coverage and (Right) Approximated coverage obtained applying Voronoi Tessellation.

CDR (Call Detail Record) databases are generated when a mobile phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS, etc.). In the process, and for invoice purposes, the information regarding the time and the BTS tower where the user was located when the call was initiated is logged, which gives an indication of the geographical position of a user at a given moment in time. Note that no information about the exact position of a user in a cell is known. From all the data contained in a CDR, our study only uses the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, the BTS tower used by the originating cell phone number and the BTS used by the destination phone number when the interaction happened.

In order to generate supervised models for the prediction of socioeconomic levels using cell phone records we need: (1) ground truth data about the socioeconomic levels; and (2) the residence location, expressed as a BTS, of the cell phone users. Given these, we will be able to compute a feature vector –for each BTS– that contains both its socioeconomic level, and the aggregated behavioral, social and mobility characteristics of the individuals that have their residence in the area of coverage of each particular BTS. These feature vectors constitute the traditional machine learning set that will be used to train and test the socioeconomic prediction models. National statistical institutes compute the socioeconomic indicators for specific geographical regions (GR) that they define. However, these GRs do not necessarily match the geographical areas produced by Voronoi tessellation, thus we first need a mechanism that assigns to each Voronoi cell (and to its BTS) a socioeconomic level. On the other hand, given that socioeconomic levels are obtained interviewing people that live within specific GRs, we need to compute the residential BTS of the individuals in our study. For that purpose, we will use an algorithm that can identify the residential BTS of an individual from its calling patterns. The following section gives more details about the data acquisition process and the mechanisms here described necessary to prepare the dataset.

3 Data Acquisition and Pre-processing

3.1 Cell Phone Traces and Behavioral Variables

For our study, we collected anonymized and encrypted CDR traces from a main city in a Latin-American country over a period of 6 months, from February 2010 to July 2010. The city, which is covered by 920 BTS towers, was specifically selected due to its diversity in socioeconomic levels. From all the individuals in the traces, only users with an average of two daily calls were considered in order to filter those individuals with insufficient information to characterize their patterns. The total number of users considered after filtering was close to 500,000. For each one of these users a total of 279 features modelled from CDR data were computed. The features include information regarding 69 behavioral variables (such as total number of calls or total number of SMSs), 192 social network features (such as in degree and out degree) and 18 mobility variables (such as number of different BTSs used and total distance traveled). Details of the most relevant variables are given in the following sections. In order to identify the residential location of each user, we applied a residential location algorithm that uses the calling patterns to identify which BTS can be defined as home. Details of the algorithm can be found in [7]. With this information, an aggregated set of features is obtained for each BTS as the average of the 279 features for the set of users for whom that BTS is their residence.

3.2 Socioeconomic Levels

The distribution of the socioeconomic levels for the city under study were obtained from the corresponding National Statistical Institute. These values are gathered through national household surveys and give an indication of the social status of a geographical region (GR) relative to the rest of GRs in the country. In our particular case, the National Statistical Institute defines three SELs (A, B, and C), with A being the highest SEL. The SEL value is obtained from the combination of 134 indicators such as the level of studies of the household members, the number of rooms in the house, the number of cell phones, land lines, or computers, combined income, occupation of the members of the household, etc. The SELs are computed for each GR defined by the National Statistical Institute which consists of an area between 1 km^2 and 4 km^2. The city under study is composed of 1,200 geographical regions (GR) as determined by the National Statistical Institute and the SEL distribution is as follows: A levels represent 12% of the GRs, B 59% and C 29%.

3.3 Matching Behavioral Variables with Socioeconomic Levels

The data described in Sections 3.1 and 3.2 provides: (1) aggregated behavioral data for each one of the 920 BTSs that cover the city and (2) a set of 1200 geographical regions (GRs) with its socioeconomic level (A, B or C). In order to create socioeconomic predictive models we need a training set that has, for

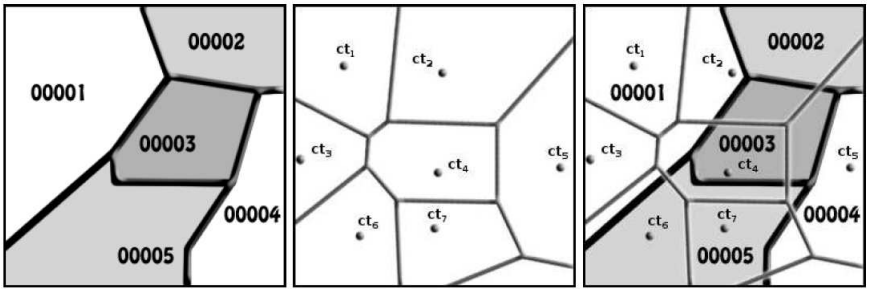

Fig. 2. (Left) Example of Geographical Regions (GR) that have a SEL associated; (Center) The same geographical areas with the BTS towers (coverage approximated with Voronoi tessellation) and (Right) The correspondence between GRs and BTS towers used by a scanning algorithm to assign a SEL to a BTS tower area.

each BTS, both its cell phone data and its socioeconomic level. However, given that the GRs do not necessarily overlap with the coverage areas, we seek to associate to the area of coverage of each BTS the set of GRs that are totally or partially included in it. Each GR within the BTS area of coverage will have a weight associated to it. The weight represents the percentage of the BTS cell covered by each GR. A graphical example is shown in Figure 2. Figure 2(left) presents the set of GRs (00001 through 0005) defined by the National Statistical Institute. Each GR has an associated SEL value (A, B or C). Figure 2(center) represents, for the same geographical area, the BTS towers (ct1 though ct7) and their cell phone coverage computed with Voronoi tessellation. Finally Figure 2(right) shows the overlap between both representations. This mapping allows to express the area of coverage of each BTS cell (ct) as a function of the GRs as follows:

$$ct_i = w_1 GR_1 + ... + w_n GR_n \tag{1}$$

where w_1 represents the fraction of GR_1 that covers the coverage area of BTS tower ct_i. Following the example in Figure 2 , ct_1 is completely included in $GR0001$ and as such $n = 1$ and $w_1 = 1$. The same reasoning applies to ct_3. A more common scenario is ct_4, which is partially covered by GR 00003, 00001 and 00005 with $n = 3$ and weights $w_1 = 0.68$, $w_2 = 0.17$ and $w_3 = 0.15$ respectively. The process to compute the mapping between the BTS coverage areas (cts) and the GRs uses a scan line algorithm to obtain the numerical representations of each GR and BTS map [8]. These representations are then used to compute the fractions of the BTS cells covered by each GR. A more detailed description of the algorithm can be found in [7]. Once each BTS tower is represented by a set of GRs and weights, we can associate a SEL value to each BTS. To do so, we first transform the discrete SEL values into a [0-100] range where values in [0-33.3] represent a C SEL, values in [33.4-66.6] a B socioeconomic level and values in [66.7-100] a socioeconomic level A. The final SEL value associated to a BTS can be obtained by computing Formula (1) assuming the central values of the

range associated with each SEL: $A = 83.3$, $B = 50$, and $C = 16.6$. Following the previous examples and assuming that the SEL of GRs 00001, 00005 and 00003 are respectively B, B and C, the SEL associated with BTS ct1 and ct3 will be 50, socioeconomic level B, while the SEL associated with BTS ct4 will be 0.68*50+0.17*16.6+0.15*50=44.3, also a B socioeconomic level.

4 Feature Selection

After the initial pre-processing, the training set consists of 920 vectors (one per BTS), each one composed of 279 features (as described in Section 3.1) with its target class, the socioeconomic level. In order to improve the prediction models, we first evaluate the features that are more relevant in our dataset. By bootstrapping the prediction models with vectors of features ordered by relevance, we expect to optimize our classification results. For that purpose, we apply two different feature selection techniques: maxrel and mRMR [9, 10]. Maxrel selects the features with the highest relevance REL to the target class, while mRMR selects the features that maximize a heuristic measure of minimal redundancy RED between features and maximal relevance REL of each feature with respect to the target class. This heuristic can be defined in two ways, as a difference (mRMR-MID) and as a quotient (mRMR-MIQ) between the relevance REL and the redundancy RED. The mRMR implementation used is available at `http://penglab.janelia.org/proj/mRMR/`. Both feature selection techniques need all dimensions to be discretized, including the target class. However, the discretization is applied only during the feature selection process. The target class is discretized as explained in the previous section: class C ranges between $0 < SEL \leq 33.3$, class B ranges between $33.3 < SEL \leq 66.7$, and finally class A ranges between $66.7 < SEL \leq 100$. The rest of the features are discretized to three values using the following scheme:

$$x^j \in (-\infty, \mu - \sigma/2) \Rightarrow x^j_{new} = -1 \quad (2)$$

$$x^j \in [\mu - \sigma/2, \mu + \sigma/2] \Rightarrow x^j_{new} = 0 \quad (3)$$

$$x^j \in (\mu + \sigma/2, \infty) \Rightarrow x^j_{new} = +1 \quad (4)$$

4.1 Top Features Selected

The three techniques used for feature selection (maxrel, mRMR-MID and mRMR -MIQ) identify a very similar set of variables as the most relevant ones. In this section we describe the top ten features after averaging their position for the three techniques used. It is important to recall that all features are computed – for each BTS – as the average of the users' features whose residence location is that particular BTS. The most relevant features 1, 2, 7 and 8 correspond to mobility variables; features 3, 5 and 9 are behavioral variables and features 4, 6 and 10 social network variables:

(1) Number of different BTS towers used (weekly): it represents the average number of different BTS towers used by an individual during the chronological period under study.

(2) Diameter of the area of influence(weekly): the area of influence of an individual is defined as the geographical area where a user spends his/her time doing his/her daily activities. It is computed as the maximum distance (in kilometers) between the set of BTS towers used to make/receive calls during the temporal period under study.

(3) Total number of weekly calls: total number of calls that an individual makes and receives every week during the period of study.

(4) Closeness of incoming SMS-contacts in relation to all communications: it is defined as the average geographical distance in kilometers of all the contacts that sent at least one text message to the individual divided by the total geographical distance for SMS, MMS and voice. Low values of this measure mean that the user's text-contacts live closer than his/her voice or MMS contacts.

(5) Percentage of incoming SMSs with respect to all incoming communications: number of received SMSs over all communications (SMS, MMS and voice).

(6) Percentage of SMS-contacts with degree of reciprocity 5: number of contacts that an individual exchanges SMS with and that account to at least five text messages per week over all the individuals' contacts (SMS, MMS and voice) that exchange communications at least five times per week during the period under study.

(7) Radius of gyration: it is defined as the root mean squared distance between the set of BTS towers and its center of masses. Each tower is weighted by the number of calls an individual makes or receives from it during the time period under study. The radius of gyration r_g and the center of masses r_{cm} are computed as:

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - r_{cm})^2}, \quad (5) \qquad r_{cm} = \frac{1}{n}\sum_{i=1}^{n} r_i. \quad (6)$$

The radius of gyration can be considered an indirect indication of the distance between home and work (and of the daily commute), given that the towers with the highest weights typically correspond to the towers that give service to the user while at work or at home.

(8) Total distance traveled(weekly): it is defined as the sum of all weekly distances traveled during the time period under study for the individuals whose residence is at that BTS.

(9) Median of total number of calls: the median of the number of calls of all the individual living in the area of coverage of a tower.

(10) Percentage of voice-contacts with degree of reciprocity 2: number of contacts that an individual exchanges voice calls with and that account to at least two calls per week over all the individuals' contacts (SMS, MMS and voice) that exchange calls at least two times per week during the period under study.

Once the features have been ordered according to their relevance, the prediction of socioeconomic levels can be formalized as a classification problem that

we solve using SVMs and Random Forest, or as a regression problem which we solve using SVMs.

5 SEL prediction as a Classification problem

The classification problem can be formalized as assigning one of the $SEL = \{A, B, C\}$ to a given BTS, and by extension to its area of coverage, based on its aggregated feature vector. Although we have tested several classification methods, we only report the results obtained by SVMs and Random Forests, which yielded the best classification rates. We have tested the classification methods with the feature vectors ordered according to each one of the three feature selection techniques described before in order to understand which one produces better results. On the other hand, we have also tested them on all of its subset vectors from 1 to 279 ordered features so as to determine the number of relevant variables needed for a good prediction rate. In all cases, the BTS dataset with the ordered features and its associated SEL was partitioned for training and testing, containing 2/3 and 1/3 respectively. The classification was implemented using the SVM library libsvm-Java [11] and the Weka Data Mining Software [12] for the Random Forest.

5.1 Support Vector Machines

SVMs have been extensively and successfully used in similar classification problems [13, 14]. We have used a Gaussian RBF kernel that is based on two parameters: C and γ. C is a soft-margin parameter that trades off between misclassification error and rigid margins and γ determines the RBF width. For each feature selection order (produced by maxrel, mRMR-MID and mRMR-MIQ), and for each subset of ordered features in $n = \{1, \ldots, 279\}$, we identify the optimum values for (C, γ) as the ones that maximize the accuracy using 5-fold cross-validation over the training set. The search was performed with values of $C \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and of $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$ [15]. Figure 3(left) shows the grid search during the cross-validation stage of one specific feature ordering.

After that, each SVM model is tested using the test set. Figure 3(right) shows the accuracy (Y axis) for each subset of ordered features (X axis) for the three feature selection techniques used. Results for datasets with more than 50 features are not shown, as the classification rate stabilizes. It can be observed that maximum relevance feature selection (maxrel) produces better accuracy results than mRMR-MIQ or mRMR-MID. The best result with maxrel is obtained when using the top 38 features (80% accuracy). A compromise solution would be using the top 17 features, given that we obtain a similar accuracy (79.1%) with considerably fewer variables. The confusion matrices when using 38 and 17 features are:

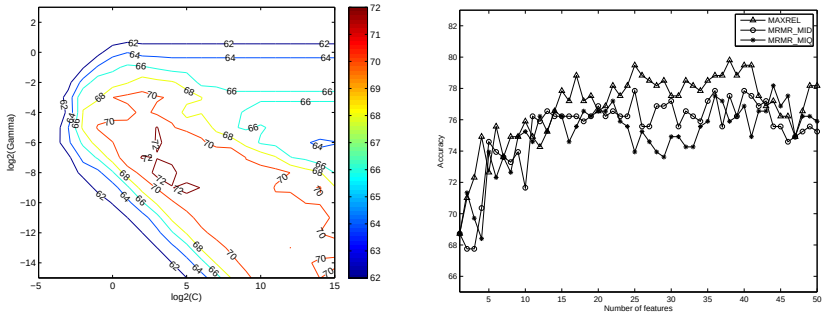

Fig. 3. (Left) Example of the identification of the optimum C and γ values when using mRMR-MIQ and the 38 most relevant features, and (Right) Correct classification rate (Y axis) for the most relevant subsets of $n = 1, ..., 50$ ordered features when using maxrel, mRMR-MID and mRMR-MIQ.

$$P^{38}_{maxrel} = \begin{pmatrix} 0.67\ 0.33\ 0.00 \\ 0.09\ 0.87\ 0.04 \\ 0.02\ 0.30\ 0.68 \end{pmatrix}, P^{17}_{maxrel} = \begin{pmatrix} 0.67\ 0.33\ 0.00 \\ 0.08\ 0.88\ 0.04 \\ 0.02\ 0.38\ 0.61 \end{pmatrix} \quad (7)$$

An interesting fact that can be observed across all confusion matrices is that if SELs A or C are misclassified, they are misclassified as B, reflecting the implicit order between the three SELs. This implies that when a classification error occurs, the closest SEL to the real one is selected, thus limiting the impact of the incorrect classification in the analysis.

5.2 Random Forest

Random Forest is an ensemble classifier in which two basic ideas are used: bootstrap sampling and random feature selection [16, 17]. Basically, Random Forest takes a bootstrap sample as the training set and the complementary as the testing set. During the training of the tree, each node and its split is calculated using only m randomly selected features, $m << M$ where M is the dimension of the feature space. We build Random Forest models with t trees where $t = \{10, \ldots, 100\}$ for each subset of ordered features in $M = \{1, \ldots, 279\}$, and for each feature selection technique used. Depending on the size of the subset M, $m = \log_2(M + 1)$ random features were considered in each split. Figure 4(left) shows the classification accuracy (Z axis) depending on the size of the forest generated (Y axis) when considering subsets of up to 50 ordered features produced by maxrel. Larger subsets did not improve classification rates. Figure 4(right) shows the maximum accuracy for each subset of features across all values of t (number of trees). We observe that the three feature selection methods reach very similar rates. The best classification rate is achieved by the mRMR-MIQ (82.4%) when using 38 features (and 44 random trees). The mRMR-MID method

reaches 80.7% with 28 trees and 41 features and maxrel yields an accuracy of 80.4% with 33 features and 83 trees.

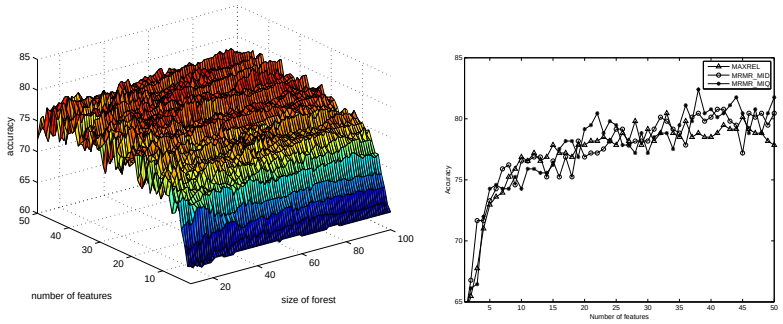

Fig. 4. (Left) Accuracy of the random trees generated for the feature subsets of up to 50 variables produced by maxrel, and (Right) Maximum accuracy obtained for each subset of features produced by each feature selection technique.

The confusion matrix of mRMR-MIQ with 38 features 8 (and most of the confusion matrices obtained) indicate that the classifier has the desirable effect of predicting the adjoining class when a classification error is made.

$$P^{38}_{mRMR-MIQ} = \begin{pmatrix} 0.77 & 0.23 & 0.00 \\ 0.07 & 0.90 & 0.03 \\ 0.02 & 0.34 & 0.64 \end{pmatrix} \tag{8}$$

6 SEL prediction as a Regression Problem

Regression techniques approximate a numerical target function by minimizing a loss function on a training set. The literature reports some cases in which the use of regression instead of classification methods improved the final prediction rates [18]. Thus, given that socioeconomic levels can be expressed as numeric intervals, we explore the computation of socioeconomic prediction models using regression. Support Vector Regression (SVR) Machines [19] are based on similar principles as SVMs for classification: the dataset is mapped to a higher dimension feature space using a nonlinear mapping and linear regression is performed in that space. An important difference between SVMs and SVRs is a loss function that defines a tube of radius ϵ around the predicted curve. Samples lying within this ϵ-tube are ignored and the model is built taking into account the remaining training dataset. The ϵ parameter needs to be determined beforehand.

Following a similar approach to Section 5.1, we use 5-fold cross validation to select the parameters (C, γ, ϵ) that minimize the mean squared error for each subset of ordered features in $M = \{1, \ldots, 279\}$ produced by each feature selection

method. We then measure the accuracy of the SVRs against the test set. Figure 5(left) shows the root mean square error (Y axis) for each subset of features (X axis) and each feature selection technique. In this case, mRMR-MID usually obtains the best results, with an RMSE in the range $(8.5, 11.5)$. However, our main interest lies not so much in the numerical socioeconomic value ([0-100]), but in the SEL class associated to that number *i.e.,* in identifying whether SEL is A, B or C. Figure 5(right) shows the accuracy results after discretizing the results of the regression from the range [0-100] onto classes {A,B,C}. Not surprisingly, the best accuracy (80.13%) is achieved when using the 38-feature subset produced by maxrel, although smaller subsets reach similar results. In our particular case, there is not a relevant improvement in the prediction accuracy when using regression as a proxy for classification. However, the use of SEL expressed numerically ([0-100]) instead of through labels, might provide more meaningful information.

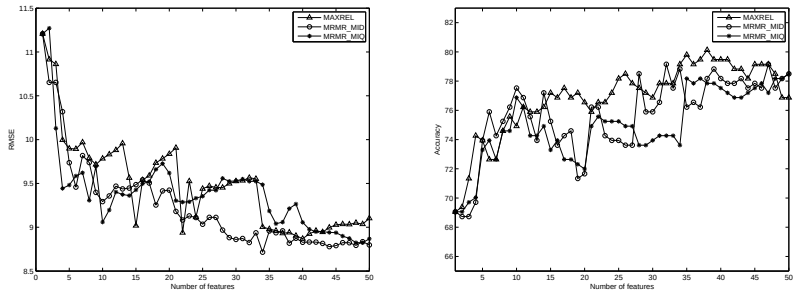

Fig. 5. (Left) Root mean squared error for each subset of features and each feature selection mechanism, and (Right) Accuracy of SEL prediction for each subset of features and each feature selection mechanism when discretizing regression results.

7 Conclusions

The identification of socioeconomic levels is a key element for both commercial and public policy applications. Traditional approaches based on interviews are costly both in terms of money and time. Thus, it becomes relevant to find complementary sources of information. Because cell phones are ubiquitously used, they have become one of the main sensors of human behaviors, and as such, they open the door to be used as proxies to study socioeconomic indicators. In this paper we have presented the use of the information collected from cell phone infrastructures to automatically assign a socioeconomic level to the area of coverage of each BTS tower using classification and regression. Each BTS tower was characterized by the aggregated behavioral, social network and mobility variables of the users whose residence lies within the BTS coverage area.

Our results indicate that call data records can be used for the identification of SELs, achieving a correct classification rate over 80% using only 38 features.

References

1. Propper, C., Diamiano, M., Leckie, G., Dixon, J.: Impact of patients' socioeconomic status on the distance travelled for hospital admission in the english national health service. Journal Health Serv. Res. Policy **12**(3) (2007) 153–159
2. Carlsson-Kanyama, A., Liden, A.: Travel patterns and environmental effects now and in the future: implications of differences in energy consumption among socio-economic groups. Ecological Economics **30**(3) (1999) 405–417
3. Rubio, A., Frias-Martinez, V., Frias-Martinez, E., Oliver, N.: Human mobility in advanced and developing economies: A comparative analysis. AAAI Spring Symposia Artificial Intelligence for Development, AI-D, Stanford, USA (2010)
4. Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Socio-economic levels and human mobility. Qual Meets Quant Workshop - QMQ 2010 at the Int. Conf.on Information & Communication Technologies and Development (ICTD) (2010)
5. Eagle, N.: Network diversity and economic development. Science **328**(5981) (2010)
6. Frias-Martinez, V., Virseda, J., A.Rubio, Frias, E.: Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. Int. Conf. on Inf. & Comm. Technologies and Development (ICTD),UK) (2010)
7. Lane, M., Carpenter, L., Whitted, T., Blinn, J.: Scan line methods for displaying parametrically defined surfaces. Communications ACM **23(1)** (1980)
8. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1226–1238
9. Ding, C.H.Q., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinformatics and Comp. Biol. **3**(2) (2005) 185–206
10. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11** (2009) 10–18
12. Burbidge, R., Buxton, B.: An introduction to support vector machines for data mining. Technical report, Computer Science Department, UCL (2001)
13. Frias-martinez, E., Chen, S.Y., Liu, X.: Survey of data mining approaches to user modeling for adaptive hypermedia. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews **36**(6) (2006) 734–749
14. Hsu, C.W., Chang, C.C., Lin, C.J.: A practical guide to support vector classification. Technical report, Department of Computer Science, Taiwan Univ. (2003)
15. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8) (1998) 832–844
16. Breiman, L.: Random forests. Machine Learning **45** (2001) 5–32
17. Frias-Martinez, E., Chen, S., Liu, X.: Automatic cognitive style identification of digital library users for personalization. Journal of the American Society for Information Science and Technology **58**(2) (2007) 237–251
18. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., Vapnik, V.: Support vector regression machines. In: NIPS. (1996) 155–161