Reality Mining

Alessandro Pomes and Domantas Meidus

November 26, 2017

Abstract

The main goal of this project is to find some correlations between the habits of some students and the type of studies thier belong. Our ideas is that we could reconize the adress of studies(or a type job for a worker) considering in which place the subjets are used to pass their time during a day. More precisely how many time a student can pass at work,at home and elsewhere
For this type of consideration team has taken the Reality Mining dataset. Below we'll provide a short description how the dataset is composed .The type of classification taken was divided in two ways: one performed with a Perceptron Neural Network and one using first Restricted Boltzmann Machines and then again a Perceptron. Our thesis is that if a generative learning algorithm is used to reconized pattern among data before the classification, then a better result in term of accuracy could be reach up.

Contents

1 Description of the problem **2**
1.1 Dataset description 2

2 Description of our approach **2**
2.1 Research 2
2.1.1 Feature Extraction 2
2.2 Classifiers 3
2.3 Validation 3

3 Implementation **3**
3.1 Data Handling 4
3.2 Missing Data 4
3.3 Feature Extraction and Selection 4
3.4 Classification 5
3.5 Main file 6

4 Results **6**

5 Conclusions **6**

1 Description of the problem

1.1 Dataset description

Reality Mining[1] was an experiment conduced from 2004-2005 at MIT Media Laboratory. It consisted to detect the the follow information of 94 subjects:

- Bluetooth devices (proximity of approximately five meters)
- call logs
- cell tower ID
- application usage
- phone status
- self-report relational data

These data were detect with the help of particular type of cellphone (Nokia 6600) where was istalled some additional software who was able to automatically run the an application called ContextLog in backround.
This application was in charge to logs all data described above. Therefore for each subject with this dataset might be extract several information during his day.It might also construct how many other people is meeting in a specific time slap. Mining the reality of almost 100 users raises justifiable concerns over privacy. However, the work in this paper is a social science experiment, conducted with human subject approval and consent of the users.

2 Description of our approach

2.1 Research

Many studies[2] [3] [4] were done for prediction of routine structure from every days people life. To reach this interesting mining from this dataset some probabilistic tool were used. Mainly Latent Dirichlet Allocation were used[3] [4]. LDA is a powerful mechanism to extract recurrent behaviors and high-level patterns with unsupervised methodology.
These researches devoted themselves to structuring the data with complete information on the student's specific positions(cell tower) and also to the information of how many people met during a day(with the help of bluetooth recording). To build the models different data structure were used, for instance paper from Ferrari, Mamei 4 used an interesting structure way where for each user are recorded several time-frames and the GSM towers where the user was connected.
Principal problem due to handle the dataset was the amount of missing data. Unfortunately when team tryed to extract some features in many cases the informations given was too partial to create a robust dataset. So we decided, inspired latter to professor and former from paper of Nathan Eagle & Alex Sandy Pentland2 to create Multidimentional-Array using the location during the days In the Section Feature Extraction is provided the description on how we organize data.

2.1.1 Feature Extraction

As mentioned above we built two type of classification problem with the same goal: provide a way to recognized the address studies from his routine.

First case of develop was to consider each subject and for each hours in a day the frequencies that a person could have been in a specific place. So for this type of solution we gathered all days information in a single matrix(for more information on the implementation consult Data Handling subsection).
For the second model development we decide to build a much deeper configuration of the data. For reasons due to the structure of RBM and to preserve information of continuous daily routine we decide to keep the structure of a week.
Inspiring from literature2 we place a single week in a long array of 840 dimension. This vector has 5 recursive structure of 168 dimension that belongs on the places we have in the dataset(Home, Elsewhere, Phone is off, No signal). For each 168 sloth that corrispond a single hour we put 1 if the student were in the place, insted of 0 it he was not(implementation detail here).
In term of clarify this structure the week obtained are random weeks divided from to type of student: The Sloan Business student and students belong to MIT. Therefore finaly we divided our dataset in 2 iphotetical weekly pattern routine, with the goal to detect it.

2.1.1.1 Feature Selection & Restricted Boltzmann Machines

On the second extension of our implementation we used a Generative learning algorithm to provide a sort of feature selection (actually this type of features mapping is not a properly a feature selection, but we are on purpose abusing this term to explain better what is our intention). To show how a eclectic could be the formulation of Neural Network we have adopted Restricted Boltzmann Machines [5] to achieve a better accuracy in classification problem. Restricted Boltzmann machine is a Unsupervised method highlight a set of latent factors. Instead of learning directly on a continuous scale, using RBM we try to discover latent factors that can explain the "activation" in a particular hour that is when a subject is a precise place; therefore if there is a precise structure in the domain data. So we were trying to collapse the information of a single week in a some much miningful features. To know this we map the original week vectors in some other trained vector that perform the follow probability:

$$p(h_j = 1|v) = \sigma(bj + \sum_i ivivw_ij)$$

we use a type of The structure

2.2 Classifiers

2.3 Validation

3 Implementation

Reality mining dataset has ~57 sub categories with each subject activity data. However it's unable to use all categories for Neural Network - only few categories data was chosen to train and test for learning. Categories that was chosen:

1. **my_affil** - this category of dataset contains each subject affilation type. All possible affilation types:
    - 'mlgrad' – Media Lab Graduate Student (not a first year)
    - '1 st yeargrad' – Media Lab First Year Graduate Student
    - 'mlfrosh' – Media Lab First Year Undergraduate Student

- 'mlstaff' – Media Lab Staff
- 'mlurop' – Media Lab Undergraduate
- 'professor' – Media Lab Professor
- sloan' – Sloan Business School.

2. **data_mat** - Inferred each subject locations at each hour of the day:
   - 1 - Home
   - 2 - Work
   - 3 - Elsewhere
   - 0 - No signal
   - NaN - Phone is off.

From data_mat dataset patterns which indicates each subject locations at each hours of the day subjects are classified to **sloan** or **no sloan**.

3.1 Data Handling

Dataset is stored in the .mat format, in order to extract this data to python data structures, the program uses scipy.io library. After all Reality Mining dataset is stored to python data structures, **my_affilation** and **data_mat** categories are extracted:
affilation = data['my_affil']
data_mat = data['data_mat']
These categories are used for Neural Network learning.

3.2 Missing Data

Reality mining dataset has a lot of missing data which is handled by excluding all missing data records in my_affil and data_mat datasets. This way of dealing with missing data prevents any error using Machine Learnings algorithm in cost of losing data. Of all missing data in my_affil and data_mat datasets 47 subjects was excluded. In result, for data preprocessing

3.3 Feature Extraction and Selection

Two Neaural Network learning algorithms are used: **Restricted Bolzman Machine(RBM)** and **Multi-layer perceptron**. For **Restricted Bolzman Machine** learning algorithm **data_mat** data will be used as features. To extract these features into required 2-nd dimensional vector, the program performs these steps:

1. Each subject activity is stored to the list if subject data_mat data complies with requirements:
   - data_mat data representing subject activity on each hours is not empty;
   - data_mat data should represent at least 7 days of subject activity.

   These conditions complies 69 subjects of 106 subjects. These 69 subjects data_mat data are stored into **features_list** for further data preprocessing.

2. Excluding NaN values from the **features_list**. Data_mat dataset contains NaN values which indicates that subject phone was off at certain hour. These NaN values are converted to Integer data type, in our program NaN values converted to value 4. Our program is not design to handle non numeric data types, so in order to perform features vector with necessary data, program converts non numeric value to Integer.

3. Creating Features vector for RBM learning algorithm. Features vector is 2-nd dimensional:

    - 1-st dimension data - Subject
    - 2-nd dimension data - Subject activity of each hour in seven days.

Features for RBM algorithm are represented in **(subject number, observation days * number of hours in a one day * all possible locations conditions)** which in real numbers are **(69, 7 * 24 * 5) = (69, 840)** shape, explanation:

- subject number - is all subjects which complies data_mat requirements.
- observation days - number of how many days subject activity is stored.
- Number of hours in a one day - since one day has 24 hours, so the number is 24.
- All possible locations conditions - All possible locations conditions are 5 (Home, Elsewhere, Phone is off, No signal).

For each subject, all 7 days of each hour data is stored for different locations conditions.
At the start features vector for each subject is filled with zeros values. Data for each locations conditions is saved in 0,1 numbers representation - all possible locations are iterated and if at the certain hour of the week subject was in that location, the value of that hour of the day changes from 0 to 1.

3.4 Classification

Since **Restricted Bolzman Machine(RBM)** algorithm is unsupervised learning, it will categorize subjects by itself by their mat_data data patterns. Data for **Multi-layer perceptron** supervised learning algorithm are classified into categories: **sloan** and **no sloan**.
For sloan category belongs all subjects who affilation type is equal: *'sloan' or 'sloan_2'*. For no sloan category belongs all subjects who affilation type is equal: *'mlgrad', '1 st yeargrad', 'mlfrosh', 'mlstaff', 'mlurop', 'professor'*.
Subjects are classified in this order:

- Total subjects: 69
- Sloans: 20
- No sloans: 49

3.5 Main file

4 Results

5 Conclusions

Notes

[1]Nathan Eagle, Alex Pentland, and David Lazer. Inferring Social Network Structure using Mobile Phone Data, Proceedings of the National Academy of Sciences (PNAS), 2009, Vol 106 (36), pp. 15274-15278

[2]Eigenbehaviors: identifying structure in routine Nathan Eagle & Alex Sandy Pentland Received: 12 September 2007 / Revised: 24 February 2009 / Accepted: 24 February 2009 / Published online: 7 April 2009 Springer-Verlag 2009

[3]Probabilistic Mining of Socio-Geographic Routines From Mobile Phone Data Katayoun Farrahi, Member, IEEE, and Daniel Gatica-Perez, Member, IEEE

[4]Classification and prediction of whereabouts patterns from the Reality Mining dataset Laura Ferrari, Marco Mamei Dipartimento di Scienze e Metodi dell'Ingegneria, University of Modena and Reggio Emilia, Italy

[5]Geoffrey Hinton 2010.August 2, 2010 UTML TR 2010–003, "A Practical Guide to Training Restricted Boltzmann Machines" Geoffrey Hinton Department of Computer Science, University of Toronto