Are you moving predictably?

Miriam Wagner, Martin Breuer, Moritz Werthebach, Timo Bergerbusch, and Walter Schikowski

RWTH Aachen, Templergraben 55, 52062 Aachen, Germany

Abstract. We analyze movements in the urban environment of the columbian city Medellín. Each movement is given as spatiotemporal pattern of with additional information about the reason, means of transportation and the corresponding person like the socio-economic status (strata), the age and gender. Since in most cases we do not have information about the actual socio-economic status of persons we firstly try different unsupervised approaches to find natural clusters. Due to bad results we introduce our preprocessing steps and switch to supervised learning. Decision trees and neural networks did neither match our performance expectations which leads to our conclusion that we need more data and information about the data in order to find a proper social stratification and to predict the given socio-economic status accurately.

Keywords: Data Mining · Clustering · Rapidminer · Cluster · Neural Nets

1 Introduction

Within the time of Industry 4.0 and various data sources the question arises, if one can define who we are by the data collected? In particular is it possible to determine the wealth of a person, only given movements of a single day? For this we considered the dataset stated in [1]. There we have a set of 124979 rows of movement data from various persons from a Columbian town, called Medellín. All the data was collected at a single day, with possibly multiple entries referring to the same person.

The data entries consist of data about the movement, like endpoints or length, and also some meta parameters like gender, age or the so-called strata of the person. The strata defines the socio-economic group, reflecting the affluence and therefore impose the ancillary costs.

Our goal is to ascertain if there is a correlation between the movements and the strata.

2 Preprocessing

In order to classify the given data into smaller test sets or mask different aspects, we have to perform some analysis.

We observe that even though we have 124979 individual lines defining a movement, there is one line defining a NotANumber-exception and therefore gets neglected for further usage.
We provide the testDataGenerator python script. Through flags and input arguments, the script is able to create all test sets considered by our clustering and neural net approaches.
We observe the following distribution over the whole dataset:

strata	1	2	3	4	5	6	Σ
abs	6963	52265	49404	8772	5536	2038	124978
%	5.57	41.82	39.53	7.02	4.43	1.63	100

We observe that there is an upper bound on equal distribution through strata 6. It has at most 2038 individual elements. Furthermore we have to make sure that two different data points, which belong to the very same person, are assigned to the same cluster. To do so we compute the value ID which identifies each person and can be used to combine movements that are considered to be from the same person, i. e. two movements correspond with the very same person, if and only if they are consecutive in the original dataset and have the same strata, age and gender. This approach is taken since the surveys are concatenated sequentially and it is unlikely, that multiple consecutive movements with same strata, age, gender belong to two different persons.

strata	1	2	3	4	5	6	Σ
abs	3153	23367	21418	3497	2083	595	54113
%	5.83	43.18	39.58	6.46	3.85	1.1	100

In Section Section 2.1 we introduce vectors representing single persons. Since strata 6 is the smallest strata with 595 persons, it limits the size of an equally distributed dataset where each data point coincides with one person.

2.1 Stratified Person Data

As stated before, instead of simple IDs for every person we expand the parsing by using a data encapsulating in a class called Person. This class stores the ID, the parameters defining a person , and all movements from that person.
Then we are able to compute the following vector, with 850 entries, for further

usage, that combines all movements of the person:

$$\underbrace{\#o_1,\dots,\#o_{413},\#d_1,\dots,\#d_{413}}_{2\cdot 413},\underbrace{AM,MD,PM,MN}_{4},\underbrace{\#r_1,\dots,\#r_7}_{7},$$
$$\underbrace{\#MoT_1,\dots,\#MoT_7}_{7},\underbrace{S_{Dest},S_{Dist},G,A,strata,strataGrouped}_{6}$$

with the following abbreviations ($1 \leq i \leq 413$, $1 \leq j \leq 7$):

o_i:	the i-th origin data point	MoT_j:	the j-th mean of transportation
d_i:	the i-th destination data point	S_{Dest}:	sum of all durations
AM:	movements at time stamp AM	S_{Dist}:	sum of all distances
MD:	movements at time stamp MD	G:	the gender
PM:	movements at time stamp PM	A:	the age
MN:	movements at time stamp MN	$strata$:	the strata (used for comparison)
r_j:	the j-th reason	$strataGrouped$:	the aggregated stratas

3 Predicting

3.1 Distance Measures

TODO (1 page)

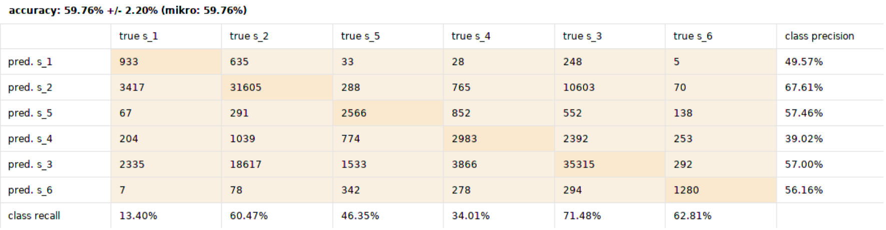
3.2 Classification

TODO (5 pages)

3.3 Neural Net

For all the neural net computations we considered person vector data sets of different sizes (c.f. Section 2.1).

We do this, because results on the normal datasets had an unacceptable performance, since only single movements and not complete paths of individuals are considered. An example training and performance measure is given in Figure 1 where unprocessed data is used. The performance is measured using 10-fold cross validation, i. e. the data is split into 10 subsets where in each iteration exactly one data set is used as test set and the other 9 as training set. The average value of the accuracy values lead to the total accuracy value of the neural net.



accuracy: 59.76% +/- 2.20% (mikro: 59.76%)

	true s_1	true s_2	true s_5	true s_4	true s_3	true s_6	class precision
pred. s_1	933	635	33	28	248	5	49.57%
pred. s_2	3417	31605	288	765	10603	70	67.61%
pred. s_5	67	291	2566	852	552	138	57.46%
pred. s_4	204	1039	774	2983	2392	253	39.02%
pred. s_3	2335	18617	1533	3866	35315	292	57.00%
pred. s_6	7	78	342	278	294	1280	56.16%
class recall	13.40%	60.47%	46.35%	34.01%	71.48%	62.81%	

Fig. 1: An example of a neural net trained without person vector data.

In the following we consider 3 neural nets $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$, all having 4 hidden layers, 50 epochs and 10 iterations. As an example of other strata aggregation we combine the stratas 1–2, 3–4 and 5–6 together and call them $\mathcal{N}_i^\star$, for $i \in \{5, 10, 20\}$. This builds a superset of the original stratas and since the stratas themselves are logically connected this task should be easier to fulfill.

For each neural net we are using equally distributed data sets with 100, 200 and the maximal amount of 595 individuals per strata which are provided by the `testDataGenerator` from Section 2. For every neural net and every set size we performed 5 independent runs and calculated the average over those accuracy values in order to have a sophisticated, comparable statements.

	#		Set size		
Name	Neurons	AG	100	200	595
$\mathcal{N}_5$	5	✗	60.03	59.92	60.18
$\mathcal{N}_5^\star$	5	✓	87.6	89.7	71.05
$\mathcal{N}_{10}$	10	✗	75.83	73.54	69.56
$\mathcal{N}_{10}^\star$	10	✓	92.93	93.48	74.58
$\mathcal{N}_{20}$	20	✗	75.45	71.14	61.87
$\mathcal{N}_{20}^\star$	20	✓	92.87	94.4	78.32

Fig. 2: The accuracy values of the neural nets. (See excel-spreadsheet)

The size of larger nets in terms of neurons is counter-productive, since if we take 50 neurons per layer we have $14 \cdot 50^4 \cdot 6 \approxeq 525.000.000$ synapses for which the input data set would be too small to have sufficient training.

4 Observations

As an overall result we get, that we can not determine the strata based on the information we have. Using various datasets (c.f. Section 2), we witnessed that having equally distributed datasets lead to an overall higher accuracy. This rules

out the bias observed in the original data, where strata 2 and 3 are very dominant (c.f. Section 2), but also reduces the set size from originally 124978 to $6 * 2038 = 12228$ entries equally distributed over all stratas.
During clustering we observed, that using a different distance measure formula would not lead to a huge difference. Also, we detected that reducing the numbers of clusters leads to better results, which reduces the significance of result that could be stated.
Using a different neural net architecture will most likely not chance the accuracy value drastically. As stated, more complex nets need more training data, which is restricted as mentioned above.
What we can observe is, that using the stratified vectors, we are able to increase the performance and accuracy, but still have no sufficient predictions. Therefore we see that the more data we have the higher the variance within the stratas themselves is. We cannot draw some kind of lines to distinguish between different entries. Datapoints, which were outliers in the smaller sets, are now not considered to be outliers, because of large numbers having the same characteristics. So the lines, we were able to draw for smaller datasets, are blurring and create some kind of transition phase.

5 Discussion

TODO (2 page)

References

1. Lotero, Laura, et al. "Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes." Royal Society open science 3.10 (2016): 150654.
2. Hudson, Rex A. Colombia: A country study. Government Printing Office, 2010.