Recursive Clustering of Mobile Phone Call Database

Kishore Rathinavel. Guided by: Prof. Pawan Lingras

Department of Electrical Engineering, Indian Institute of Technology Gandhinagar

ABSTRACT

Phone call records of a group of customers can be a great source of information for user modeling with a view to glean customer behavior and preferences. Customer profiling is one of the preliminary and primary aspects of a customer relationship management strategy. The calling pattern can be useful in suggesting appropriate modifications to the customers calling plan. Extending the analysis for all the users can be useful for the phone companies to better manage their resources as well as to maximize utilization and profits. This project builds on earlier studies of mobile phone data mining to explore a novel technique of reiterative network mining. Using the clustering techniques we explore the idea of applying the techniques recursively. The idea is that after each clustering, the data is modified slightly to retain the information about the clusters and further clustered repeatedly. We expect to see new patterns which were previously hidden and thereby advance the research in this field.

INTRODUCTION

DATA SET:

A Phone Call database during the year 2004-2005 was utilized in order to accomplish this preliminary clustering project. The dataset includes detailed logs of more than a 100 users' mobile phone activity over a nine month period. Since only voice calls and SMS are being considered the data set will be composed of phone logs of 92 users.

INITIAL CLUSTERING:

The phone logs are initially inspected by clustering the phone calls. The presence of outliers is observed, mainly with respect to duration of calls. Hence the data set is divided into two parts- one with high duration calls and the other low duration calls.

OPTIMUM NUMBER OF CLUSTERS:

The average values of each parameter for each phone number is obtained and a new dataset is formed out of this. The optimal number of clusters of this new dataset by observing parameters like withinss, betweenss the Davies-Bouldin index. The optimum number of clusters was found to be 4.

DESIGN OF EXPERIMENT

Step I: Perform initial clustering analysis on the Numbers table which contains data of each phone number and obtain the optimum number of clusters say ‘k’.

Step II: Add ‘k’ columns to the table which contains data of Numbers Table.

Step III: Each origin number has many destination numbers in the Calls table. Using this table we find the number of destination numbers for each origin number in each of ‘k’ clusters. This data is entered into the additional columns of the Numbers table.

Step IV: Perform the next iteration of clustering by using the new data set we have prepared. Repeat the process from step II till the clustering results converge.

GRAPHICAL REPRESENTATION

MATHEMATICAL FORMULATION

Variable description for the equations:

pn - phone number data D - Data set
s - static data i - iteration number
d - dynamic data j – phone number

$$pn_j^i = (s_j^i, d_j^i) = (s_j^0, d_j^i)$$

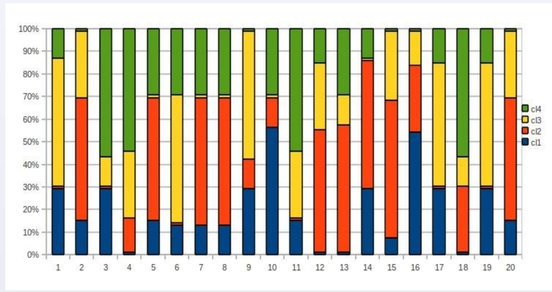
$$s_j^i = s_j^{i-1} = s_j^0$$

$$d_j^i = (m_{j1}^{i-1}, m_{j2}^{i-1}, \ldots, m_{jk}^{i-1})$$

$$D^i = (pn_1^i, pn_2^i, \ldots, pn_{j\max}^i)'$$

STATIC RESULTS

Trends in the cluster sizes from iteration 1 to 20



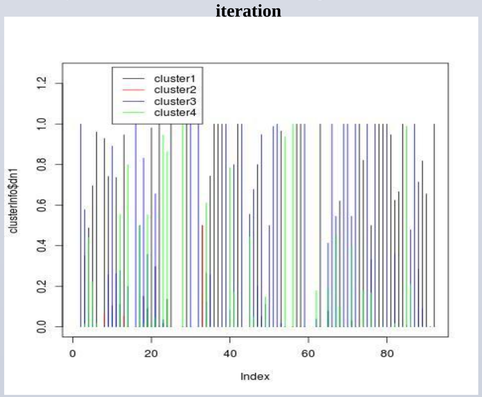
We draw cluster profiles from the cluster centers. The profiles are summarized here:

Parameter	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Number of calls:	Lowest	Low	Highest	Low
Average Duration:	Lowest	Highest	Moderate	Moderate
Weekend calls:	Highest	Moderate	Low	Lowest
Day-time calls:	High	Moderate	Moderate	High
Outgoing calls:	Lowest	High	Moderate	High
Missed calls:	0	High	Moderate	Moderate
SMS calls:	Highest	Lowest	High	Moderate
Voice calls:	Lowest	High	Low	High
Long duration calls:	0	Highest	Low	Low

DYNAMIC RESULTS

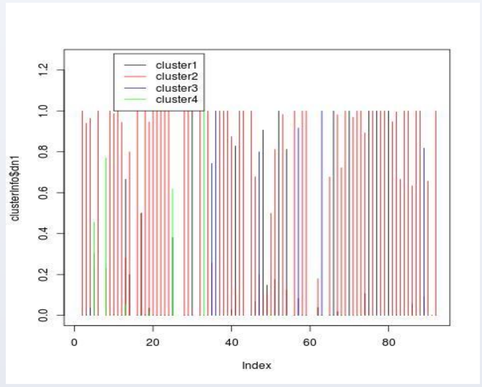
The normalization is such that m_j^i give a measure of the probability of the destination numbers belonging to a particular cluster (1, 2, 3 or 4) for each of our 92 phone numbers. The colored lines indicate the probability of a phone number's calls to belong to a particular cluster number from among clusters 1, 2, 3 and 4. Here we show the plot for the 1st and the 2000th clustering.

Probability of each Phone number to belong to a cluster after the 1st iteration



DYNAMIC RESULTS(cont…)

Probability of each Phone number to belong to a cluster after the 2000th iteration



OBSERVATIONS:

1. Some of the numbers have high probability of their destination numbers to be in a particular cluster where as some of the numbers have a moderate probability.
2. The numbers with high probability of their destination numbers to belong to a particular cluster retain their high probability values through the iterations but the phone numbers with low probability values fluctuate a lot.
3. Overall the basic structure of the graph does not change much over the iterations except for most of the probability values becoming higher marginally.

Although we notice that iteration to iteration the phone numbers belong to different clusters as well as the cluster centers change, we can derive certain very dominant trends which are visible even when the number of iterations is as large as 2000.

Cluster centers for the dynamic section of the clustering

Cluster Number	$m_{j,1}^i$	$m_{j,2}^i$	$m_{j,3}^i$	$m_{j,4}^i$
1	0.000	0.000	0.000	0.000
2	0.286	0.381	0.000	0.166
3	0.454	0.310	0.000	0.116
4	0.290	0.351	0.012	0.152

Based on the above table we can infer:

- Cluster 1 is least socially connected
- Cluster 3 is least contacted by other clusters
- Cluster 3 is mostly in contact with cluster 1 and 2
- Cluster 4 is most socially active and has connections in every cluster
- For each cluster, cluster 2 phone numbers are contacted the most
- For each cluster, cluster 4 phone numbers are contacted the least.