Using Data Mining for Mobile Communication Clustering and Characterization

A. Bascacov*, C. Cernazanu** and M. Marcu**
* Lasting Software, Timisoara, Romania
** Politehnica University of Timisoara/Computer and Software Engineering Department, Timisoara, Romania
adela.bascacov@lasting.ro, cosmin.cernazanu@cs.upt.ro, marius.marcu@cs.upt.ro

***Abstract*—Nowadays, most telecommunication companies use data mining algorithms to analyze and profile their customers based on their communication behaviour. However, similar analysis can be elaborated on call data available to any organization that use any type of communication technologies (e.g. mobile, PBX, VoIP, teleconference). Therefore, in this paper we investigate how data mining algorithms can be used to characterize employers' communication patterns and their influence on business related activities.**

I. INTRODUCTION

Both IP multimedia communication and telecommunication industries generate and store huge amount of high-quality data, which can be used to extract relevant information related to business specific parameters. Due to the size of data, a human analysis is not possible. These data are classified as: call detail data, network data and customer data [1]. Nowadays, telecommunication companies utilize data mining to improve their marketing efforts, identify fraud and optimize their network equipment infrastructure [2].

Data mining techniques are successfully applied for finding hidden correlations, patterns and unexpected trends in very large datasets. The problem of revealing users calling network profile and statistical calling information based on their calls it is essential in various applications. Data mining is the process of analyzing and automatically extracting interesting and previously unknown dependencies and relationships among data, leading to a better understanding. Data mining use defined methods for analyzing current and historical data in order to predict future trends.

Data mining uses a combination of techniques from machine learning, statistics and database technologies. The most commonly used techniques in data mining are: artificial neural networks, decision trees, genetic algorithm, rule induction and the nearest-neighbor method. Each of these techniques analyzes data in different ways.

The data mining process usually consist of: job understanding, data understanding, data preparation process modeling, process evaluation and deployment tasks [3]. The data cannot be used directly, it has to be prepared. In order for the hidden usage patterns to be discovered, data processing is required. The most import decision is related to choosing the summary features (variables). Choosing the correct features lead to an efficient description

The clustering model is a more suitable data mining approach to find natural groupings within mobile customers, based on their mobile usage behavior [4]. The [4] paper demonstrates how Self-Organizing Maps, an unsupervised learning neural network paradigm, can be used for clustering mobile customers based on the call detail record. The major clustering techniques are classified into partitioning methods, hierarchical methods and density - based methods. Several algorithms are used for clustering purposes. The algorithms can be based on supervised or on unsupervised learning. Supervised learning uses one or more manually clustered data training sets in order to assign new data set members to clusters. Unsupervised learning finds the underlying patterns from the data set autonomously and proposes these as clusters [4].

Data mining process is used in a large variety of activities in business, science and engineering to make better business decisions and strategies, by discovering patterns and relationships in the database. Most of the research in this domain shows that in the telecommunication industry the data mining techniques can be used with success in application like: market management, fraud detection and customer profiling.

II. DATA MINING ON TELEPHONY DATA

The telecommunications industry is an industry that generates a lot of data due to the large number of phone calls that are made every second. For this reason the use of data mining techniques in the industry has beneficial effects both in improving services and in providing important information for the operator.

There are a large number of data mining algorithms that can be selected to process these data (k-means, Random Forest, logistic regression, etc.). Nevertheless, the fastest and commonly chosen for data processing algorithm is the k-means algorithm.

K-Means algorithm is a heuristic algorithm and therefore it does not guarantee finding the global optimum. This issue keeps the phase of initialization of clusters. Because it is a fast algorithm, it is recommended to run the algorithm several times and in the end the user must choose only one solution.

A pseudo-code for this algorithm can be observed below.

1. Step 0 Let O be a set of observations. An observation is a vector composed of several components (metrics telephone).

2. Let C(n) be a set of clusters. The number n must be provided by user and represents the number of clusters needed to process these data.
3. Step 1 Initialization: Choose random n observations from entire set O and initialize the cluster members to these observations.
4. Step 2 Assign each observation from O set to the nearest cluster. The distance between a cluster and an observation can be a Euclidian distance, a Manhattan distance or another distance.
5. Step 3 Recomputed the entire cluster set C(n) based on their subset of observation.
6. Step 4 If there are any changes on members from cluster set, go to Step 2
7. Final StepShow the structure of the clusters and classify the observations for each cluster.

The differences to the classical algorithm arise from the Step 1 and Step 2. Thus, for a faster convergence, the clusters can be chosen to be uniformly distributed in observation space. It is also common practice to initialize clusters with members that are most representative of the entire set of observations.

The Euclidean distance was chosen to run the algorithm. For a better representation, all the data was scaled in the range [0,1]. To do this scaling, for each of the components that compose an observation, we determine the maximum value.

For an efficient running of the algorithm, the components that make up observations should not be redundant. It is difficult to identify the data redundancy, because it requires additional processing that can result in loss of important information. One possible solution to this would be to apply an algorithm to identify important components (e.g. PCA - Principal Component Analysis), but this is not the part of this article.

We mention that the main point of the article is finding available relationships between users and that is why we wanted to analyze the initial data (gross - without alteration) with no changes. For this purpose, an observation is made up of all the data collected from the mobile operator.

III. Experimental Setup and Data Inputs

The IP multimedia and telecommunication industry generates and stores a high amount of data, therefore a manual analysis is not possible. Call detail records include descriptive information about each call made in a telecommunication network. Data mining algorithms take as source of information Call Detail Records (CDR's) extracted during calls.

To uncover usage patterns, the information extracted requires data processing. The CDR reveals information at the level of individual phone calls. In data mining applications the aim is to extract knowledge at the customer level. One solution is to summarize the call detail records associated with a user into a single record that describes the user calling behaviour [2]. The users' records (the attributes) are given as inputs to the clustering algorithm.

Data sets used for this analysis contain information describing the calling behaviour of a user. Two sets of CDR data were drawn during a period of one year. Collected data was saved in a table that contains the entire year's records. Additionally, other tables contain customer information for each individual month. The first data set contains user records characterized by 24 attributes, while the second one by 20 attributes. The measures chosen to represent the user attributes in our analysis are described below.

Below is the list of attributes included in both data sets:

1. Number of calls for different services used (voice, SMS, data): describe the frequency of usage of each type of call. Calls are restricted according to the direction of call: we will take into consideration only outgoing calls. The total duration and the total number of calls describe the user's activity level.

- **voice_out_calls** – number of outgoing voice calls;
- **sms_out_calls** – number of outgoing SMS calls;
- **data_out_calls** – number of outgoing data calls;

2. Total Duration of calls:

- **voice_out_duration** – duration of outgoing voice calls;
- **data_out_size** – size of -outgoing data calls;

3. Calling Behaviour by Day/Time of the day for different call types: describes the calling behaviour of a user during a workday or weekend for different services used

- **weekend_voice_out_calls –** number of outgoing voice calls during weekends;
- **weekend_sms_out_calls** – number of outgoing SMS calls during weekend days;
- **weekend_data_out_calls** – number of outgoing data calls during weekend days;
- **workdays_voice_out_calls** – number of outgoing voice calls on working days;
- **workdays_sms_out_calls** – number of outgoing SMS calls on working days;
- **workdays_data_out_calls** – number of outgoing data calls on working days;
- **worktime_voice_out_calls** – duration of outgoing voice calls on working hours (8:00 a.m. – 6:00 p.m.);
- worktime_sms_out_calls (duration of outgoing voice calls on working hours (8:00 a.m. – 6:00 p.m.))
- **worktime_data_out_calls** – duration of outgoing voice calls on working hours (8:00 a.m. – 6:00 p.m.);
- **offtime_voice_out_calls** – duration of voice calls made outside office hours;
- **offtime_sms_out_calls** – duration of SMS calls made outside office hours;
- **offtime_data_out_calls –** duration of data calls made outside office hours;

Attributes characteristic only to the first data set describe the communication realized outside the network:

4. Outside the network communication or different call types: total number of calls for mobile, fixed, international or roaming calls are added as attributes in the data set.

- **roaming_voice_out_calls** – number of outgoing roaming voice calls;
- **roaming_sms_out_calls** – number of outgoing roaming SMS calls;
- **roaming_data_out_calls** – number of outgoing roaming data calls;
- **internat_voice_out_calls –** number of outgoing internet voice calls;
- **internat_sms_out_calls** – number of outgoing internet SMS calls;
- **mobile_voice_out_calls** – number of outgoing mobile voice calls;
- **fixed_voice_out_calls** – number of outgoing fixed voice calls;

The second data set are contains additional attributes that describe the calls realized for business purposes:

5. Business purpose calls: this information is extracted based on contacts which communicate with the user.

- **business_voice_out_calls** – number of outgoing voice calls made for business purposes only;
- **business_sms_out_calls** – number of outgoing SMS calls made for business purposes;
- **business_data_out_calls** – number of outgoing data calls made for business purposes.

Users' records will be used as inputs for the clustering algorithm.

The K-means algorithm involves grouping users into clusters and it is executed in two steps:

1. First, in order to find the centroids of the table which is provided as data input, we have to select a table, as well as the number of centroids we wish to extract. The algorithm will display the structure of the – obtained centroids.

2. The second step involves determining to which cluster the users belong to. A table is selected, and based on the centroids structure generated in the first step, the users of that table are added to the appropriate cluster.

For both data sets there is one table which contains user records for the entire year (yearly table), and 12 tables for each month (monthly tables), describing the user communication made during that particular month. Based on data in these tables, we conducted several tests using various methods.

a) In the first analysis we calculated the structure of the centroids for the table which contains records of the entire year. Then, we applied this centroids structure on each of the tables that contain monthly data. The result is a users-clusters correlation per month.

The centroids structure is also applied separately to the yearly table in order to find, at a year level, to which cluster the user belongs to.

b) In the second analysis we calculated the structure of the centroids for each monthly table and then applied the structure on that same table in order to find the users-clusters correlation. Results are compared with the ones generated in the first analysis.

c) A third analysis consists of creating a new table which contains records from all the monthly tables. The structure of the centroids is calculated based on this table and then applied to the other tables (the yearly table and monthly ones).

For each of the method proposed we –repeat the analysis above,–varying / alternating the number of centroids to 4, 6, and 8.The number of centroids represents the number of clusters in which the users will be grouped after running the algorithm.

IV. Experimental Results

The proposed experiments have been applied on one year telephony records of a local company working in the computers and software industry. The analyzed telephony database stores more than 25,000 call detail records per month for around 150 employees. The database has been created importing the CDRs provided by telephony service providers through electronic invoices. The imported CDRs have been preprocessed using a call accounting application called UniTEL (http://unitel.lasting.ro/).

A. *Clusters stability*

A number of 20 (test set 2) or 24 (test set 1) attributes have been extracted from the database for every employee. The extracted test sets have been grouped into 4, 6 or 8 clusters of communication models, using one year data set. Next, each month's data set has been distributed into the available clusters using the minimum distance. For 6 clusters the one year data set for several users is presented in Fig. 1. Every cluster is represented by a number and a color. Similar colors represent similar clusters.

User	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
U304	2	5	2	2	5	2	5	2	2	2	2	2	3
U430	4	4	6	6	6	4	4	4	4	6	6	4	1
U268	4	6	6	6	4	4	6	4	1	4	4	4	1
U157	4	4	4	6	4	4	4	4	4	4	4	4	1
U421	4	4	6	4	4	4	4	4	1	4	6	6	4
U425	3	3	5	6	3	3	3	3	3	3	3	5	4
U002	5	5	5	5	5	5	5	5	5	5	5	5	6
U034	1	1	1	1	1	1	1	1	1	1	1	1	1
U004	6	6	6	6	6	6	4	6	6	6	6	6	4
U436	6	5	6	6	6	6	6	6	6	6	5	6	4
U171	5	5	5	5	5	6	5	4	5	5	5	5	3
U125	5	5	5	5	5	5	5	3	3	3	5	5	4
U290	1	1	1	1	1	1	1	1	1	1	1	1	1
U005	3	3	3	3	3	3	3	3	3	5	3	3	3
U254	6	4	6	6	6	6	6	6	6	6	6	6	1
U107	4	6	6	6	6	4	4	4	4	4	4	4	1
U010	5	6	5	6	6	5	5	5	5	5	5	6	4
U033	2	2	2	2	2	2	2	2	2	2	2	2	3
U169	4	4	4	6	4	4	4	4	4	4	4	5	1
U003	5	3	2	5	5	5	5	5	5	5	5	5	5
U271	5	5	5	5	5	5	4	6	6	6	5	5	4
U044	2	2	2	2	1	1	2	2	2	2	2	2	5

Figure 1. One year clusters' stability analysis

In Fig. 1 similar clusters are observed for every user during the year. That means a user has the same communication behaviour every month in the year. Some

exceptions are observed for almost all users, exceptions that can be correlated with changes in their communication behaviour (e.g. vacancy periods or peak business activities).

B. *Clusters analysis*

Next, we analyzed the clusters and the relation between attributes inside clusters (Fig. 2). This analysis allows us to identify similar characteristics of employees in every cluster. Aspects like preferred communication channel (voice, SMS, data), length of communication (duration, size), time of communication (day hours) have been investigated.

Clusters	voice_out_calls	sms_out_calls	data_out_calls	voice_out_duration	data_out_size
1	35	5.095238095	9.761904762	4368.666667	11496902.95
2	340.3333333	131.3333333	1256.666667	54337.33333	386944881.3
3	125.8888889	27.22222222	330.7777778	14784.77778	294844823.3
4	102.8095238	13.71428571	12.66666667	12247.28571	5476750.19
5	246.5333333	53.53333333	132.4	30790.53333	61552110.67
6	173.6666667	17.83333333	12.5	20397.66667	2191544.667

Figure 2. Clusters' centroids

The following graphics plot clusters by various couples of attributes. Each circle represents a cluster and its size represents the number of elements contained by it. In Fig. 3, clusters are presented by their number of voice calls and SMS count. It can be observed a direct relation between number of calls and number of messages for clusters 1, 4, 5 and 2. But employees in cluster 3 prefer SMS communication unlike the employees in cluster 6 which prefer voice communication.

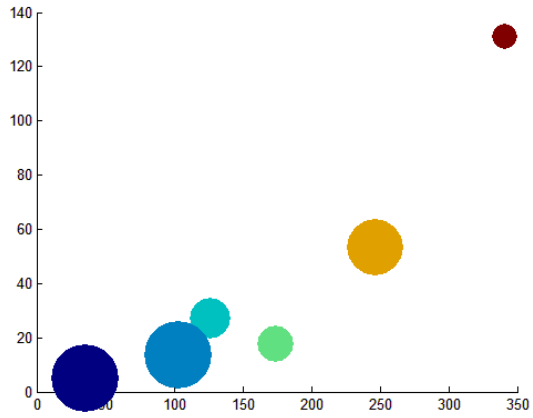

Figure 3. Voice out calls vs. SMS sends

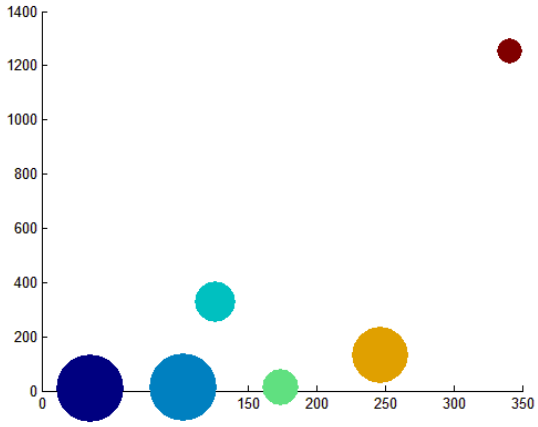

Figure 4. Voice out calls vs. initiated data traffic

In Fig. 4 the clusters are plotted by their voice and data attributes. Employees in clusters 1, 4, 6 and 5 do not use or use less data communication plans, compared with the ones in 3 and 2 which are more data oriented. The graphic in Fig. 5 shows the clusters by their messaging and data traffic attributes.

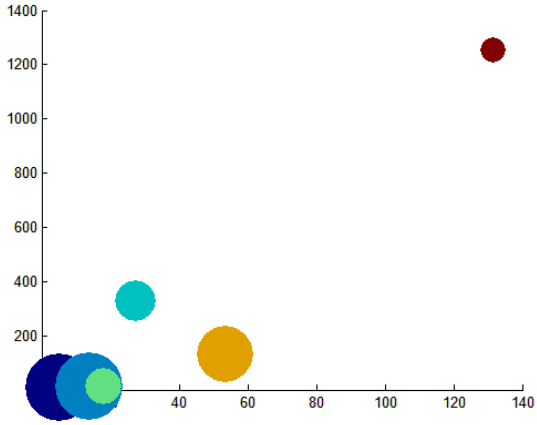

Figure 5. SMS traffic vs. initiated data traffic

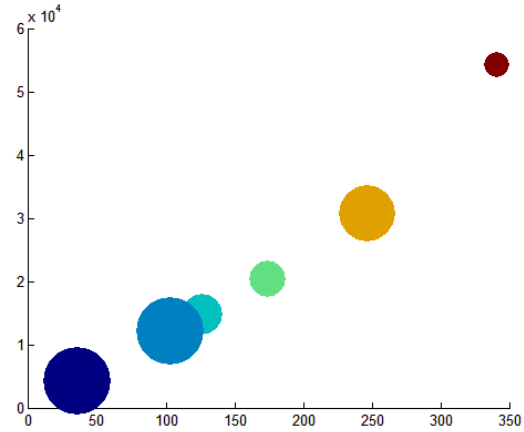

Figure 6. Voice out calls vs. average call duration

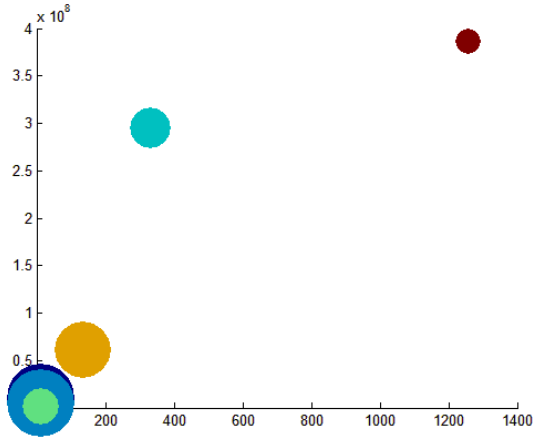

Figure 7. Initiated data traffic vs. traffic size

C. *Business analysis*

In the end we tried to identify relations between employees' communication behaviour and their business position or activity. We evaluated the job description of the employees in every cluster. In this analysis we found

similar positions of the employees inside clusters. For example in Fig. 2, cluster 2 (red) contains employees in top management positions. Cluster 5 (orange) is formed mainly by people that work in sales. Finally, cluster 1 (cyan) contains the less communicative people, the technical ones.

V. CONCLUSIONS

In this paper we investigated how data mining algorithms can be used to characterize employers' communication patterns and their influence on business related activities.

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/89/1.5/S/57649, Project ID 57649 (PERFORM-ERA), co-financed by the European Social Fund – Investing in People, within the Sectorial Operational Programme Human Resources Development 2007-2013.

REFERENCES

[1] I. O. Folasade, "Computational Intelligence in Data Mining and Prospects in Telecommunication Industry", *Journal of Emerging Trends in Engineering and Applied Sciences*, vol. 2, no. 4, pp. 601–6051, Aug. 2011.

[2] G. M. Weiss, "Data Mining in the Telecommunications Industry", *IGI Global – Section: Service*, 2009.

[3] J. K. Pal, "Usefulness and applications of data mining in extracting information from different perspective", *Annals of Library and Information Studies*, Vol. 58, pp. 7-16, Mar. 2011.

[4] K. Tulankar, M. Kshirsagar, R. Wajgi, "Clustering Telecom Customers using Emergent Self Organizing Maps for Business Profitability", *International Journal of Computer Science and Technology*, Vol. 3, Iss. 1, pp. 256-259, Mar. 2012.