U.P.B. Sci. Bull., Series C, Vol. 79, Iss. 2, 2017 ISSN 2286-3540

EMPLOYING USAGE DATA ANALYTICS FOR OBTAINING TELEMEDICINE INSIGHTS

Daniela Anca SARBU[1]

Over the past years, telemedicine strengthens its position as a key player in the health care industry transformation. Most of the studies and researches in the telemedicine area were focused on improving the delivery of existing services (by using cutting edge technology, new devices, Machine-to-Machine (M2M), creating the related legislation, etc.) or developing new services. This paper explores a new line of research (the analysis of data generated from using telemedicine services) and provides a description of the proposed approach for collecting, analyzing and interpreting telemedicine usage data. The chosen approach employs traditional methods by using data mining algorithms and spatial data analysis for predicting indicators such as user`s mobility.

Keywords: telemedicine, data mining, spatial data analysis

1. Introduction

Telemedicine is considered to be a suggestive example of frugal innovation, as it has been discovered out of the need of overcoming distance barriers and delivering health care services at a distance, to less accessible geographical areas as described in [1].

The usage of telemedicine was initially intended for rural communities, spanning across less accessible geographical areas to improve the quality of health care for those who live in remote or isolated areas where access to quality health care has traditionally been a problem [2]. However, the benefits it could bring in terms of time saving (by not needing to go to a clinic and get tested for receiving a diagnostic and having a receipt issued, etc.), travel avoidance (by being remotely monitored and eliminating the need for patients to be at the same location as the health care provider), by-passing language barriers (being able to travel and still receive medical assistance in your own language) and making available the best professional (specialized surgeons can perform surgery on a patient even though they are not physically in the same location) has made telemedicine a big focus in the urban areas as well. Telemedicine has known a big evolution when it comes to the type of services offered going from general health care delivery to tele-surgery (remote surgery using a robotic tele-operator system controlled by the surgeon,

[1] PhD Student, Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, Romania, e-mail: sarbuanca@gmail.com

where the remote operator may give tactile feedback to the user; remote surgery combines elements of robotics and high-speed data connections), tele-pathology (pathology practiced at a distance), tele-radiology (sending radiographic images from one location to another), etc.

Technologies that became part of standard practice have been and are a result of considerations of the technology costs, gained benefits and ease of implementation as also stated in [3]. Telemedicine has started to become an important player in a multitude of medical treatments and practices. As the services have grown in popularity and are more intensively used, terrabytes of usage data is generated, and this is not only patient`s information (although of course it is one aspect) but also solution usage data. The usage data is particularly interesting in this context where geo-localization and resource optimization are key words for telemedicine concept. Structuring and analyzing this unexplored data is of great value.

The focus of the study is a subject that has not been explored yet: a solution designed to provide telemedicine usage data analysis and its correlation with the geographical area where the telemedicine services were accessed.

In the previous studies presented in [4] and [13], the emphasis is put on the description of the architecture for the proposed analytical platform and results visualization. The solution design is based on a datamart (modeled with snowflake schema) that aliments an online analytical processing (OLAP) cube, which will be queried based on the different areas of interest. Results visualization is done in an interactive manner, through personalized dashboards that contain the aggregated information at the highest level, enabling decision making in telemedicine services offering and delivery.

This paper, however, tackles more the area of data collection and correlation, while [4] contains a more detailed description of the chosen fields to enter the study when it comes to Call Detail Records (CDR) data. Besides the depiction of the solution’s data sources, another focus is presenting an example of how data mining algorithms can be employed in predicting different usage indicators such as user`s mobility while accessing the services and generating traffic. All this with the aim of obtaining a better understanding of usage trends and patients mobility profile, having as an ultimate purpose insides gaining on possible services improvements or facilitating decisions that would help overcome a difficult context. This way structuring and analyzing telemedicine usage unexplored data is of great value for service offering and optimization.

2. Solution`s data sources - collection and correlation

One of the biggest challenges of the research was identifying the data that should enter the study and how it should be best correlated so that it depicts business reality. As underlined in [14], “Business understanding”, “Data understanding” and “Data preparation” are the data mining phases of Cross Industry Standard Process for Data Mining (CRISP-DM), that represent a starting point for creating the data mining structure.

One crucial data source for this analysis is represented by CDRs. This type of data is stored by telecommunication companies for billing purposes and has a high degree of confidentiality and security, as they expose sensitive information about subscribers (the detailed description of their usage of telecommunication services). A walkthrough how the data is generated and the list of fields together with their functional description can be found in an earlier paper [4].

This type of data is provided either in already aggregated datasets or has been provided in the past for research purposes as part of data mining contests organized by Orange (“Data for Development” challenge) or Telecom Italia. As stated in [5] for the CDR sample data that was analyzed during the Orange contest, the customer identifiers were anonymized by Orange Ivory Coast and all data processing was done by Orange Labs in Paris. CDRs used had the following standard format: timestamp, caller id, called id, call duration, antenna code. There is however a tool for generating CDR sample data online, GEDIS Studio (http://www.gedis-studio.com/online-call-detail-records-cdr-generator.html). GEDIS Studio is a generator for test data, available online, that produces realistic datasets by combining more than 30 generation rules. The input for the CDR generating solution is an XML file containing all the configurations for the usage profiles that need to be illustrated.

For obtaining the CDR sample data that enters this study, the GEDIS configuration files were modified accordingly to simulate the activity of a set of customers (with similar conventions for subscriber identifiers as for Orange Romania) during a period of time (first five months of 2015) with different usage patterns (voice and data traffic distribution). Based on this XML configuration file, CSV output files were created with the corresponding CDRs.

On the other hand, when it comes to localization of mobile subscribers, in [6], tree distinct main entities are identified as having access to such information, besides the mobile subscriber himself. The service provider holds all the location data of its users. This information can be accessed also by law enforcements agencies that can subpoena the information. Also as mentioned in [7], due to E911 mandate telecommunication companies are obliged to localize the mobile stations with a specific accuracy. The third category is represented by other external entities such as other users with no explicit access.

Telecommunication companies store information about their antennas. As stated on Orange site http://www.orange.ro/about/filiale-france-telecom.html, the location of base stations and other radio transmitters can be found on the website www.cartoradio.fr belonging to the National Agency of Frequencies (ANFR).

Also, the Office of Communications (Ofcom) created an Internet site: www.sitefinder.ofcom.org.uk, where you can find out the location of base stations in the United Kingdom and you can read reports on the measurements carried out by this institution. Sitefinder is hosted by Ofcom on behalf of Government and it represents a voluntary scheme under which mobile network operators make information available on the location and operating characteristics of individual base stations, so that people who wish to inform themselves about this can do so. The data within Sitefinder is owned by the mobile network operators, who supplied it on a voluntary basis. Moreover, there are dedicated sites such as http://opencellid.org/ where a data source for GSM localization can be found. This data source is continuously updated and as it is stated today on their wiki page(http://wiki.opencellid.org/wiki/What_is_OpenCellID): more than 15,000 contributors have already registered with OpenCellID, contributing more than 1 million new measurements every day in average to the OpenCellID database. Also, OpenCellID is described as a collaborative community project that collects GPS positions of cell towers, used free of charge, for a multitude of commercial and private purposes.

The data that enters in this study was downloaded on 30.06.2015. On the downloaded database, a filter was set on the MCC field to have the value 226 (corresponding to Romania) and MNC field was set-up to value 10 (corresponding to Orange telecom operator). The most important fields that will enter the study are: LAC, CellId, longitude and latitude corresponding to 10689 antennas.

However, what would be important to mention here is that for the purpose of the analysis (which includes also dashboards with maps showing the different geographical spread of the usage measurements of telemedicine services) it is needed to have more antenna attributes populated such as city, street, street no, postal code, etc. These attributes can be populated by reverse geocoding, meaning convert the pair latitude/longitude to their approximate address. There are several online tools that can provide batch reverse geocoding. The one that was used for this paper (http://www.doogal.co.uk/BatchReverseGeocoding.php) uses Google Maps services to convert the latitude and longitude to an address.

On http://www.sql-server-performance.com/ there are a series of articles written by Siddharth Mehta about working with spatial data, especially when it comes to SQL Server Integration Services (SSIS). In [8] the main topic is related specifically to geocoding text-based spatial data for use in SSIS Packages. Meaning an address is given, the aim is to obtain latitude and longitude coordinates that are vital for spatial data analysis. With this purpose, the approach used was to subscribe to Bing Maps webservices, and then create a proxy class for Geocode webservice. Afterwards, a package which reads this data was created and a call was made to the Geocode function of Geocode webservice and retrieved the latitude and longitude information returned by the webservice. This can be done using SSIS packages and in a similar manner the process of reverse geocode can be also achieved.

One other great challenge of this paper and the proposed solution is to identify the traffic (be it voice, SMS, data or M2M) generated by using telemedicine services from the regular traffic corresponding to a customer that is accessing other services than telemedicine. Also going a step further, another challenge would be to add more precision in identifying exactly what was the accessed telemedicine service that caused traffic generation: has the customer tried to access data from his personal medical record or has he requested a video consultation? What was the particular service he was using?

Addressing these challenges depends fundamentally on the choices the telecommunication operator has made for the design and implementation of the telemedicine solution. Such solutions and services can vary greatly from one telecommunication operator to another.

For the scope of this analysis it was taken as input Medic4all Orange Romania implementation, where not only M2M traffic is important to be identified, but as well voice and data traffic.

The issue of identifying M2M traffic can be addressed in two ways. Either this can be marked through the unique identifier of the devices that are taking the measurements. In this case each M2M device that is connected to the network has a dedicated International Mobile Subscriber Identity (IMSI) and as part of the solution implementation, these IMSI are flagged and known as corresponding to M2M devices.

In the case of the telemedicine solution proposed by Orange Romania for the two specific M2M services (“Vital signs measurement” and “Watch ME”) the measurements are send via internet to a telemedical monitoring center. This data traffic triggers the creations of corresponding CDR lines where the IMSI generating the traffic is identified. If the IMSI can be found in the pool of IMSIs that are dedicated to the M2M devices, then the traffic can be flagged as M2M. Another option would be to identify the correspondent to which the information is sent via data traffic. In our case, this is represented by the web medical record. If the corresponding site is dedicated to M2M related services, then that traffic can be flagged as M2M. This latter approach is the one chosen for the purpose of this paper.

The major limitation of this study that could affect the solution`s success is the access to data, more specifically having access to real Call Detail Records data. Because of confidentiality and security clauses that telecommunication companies have in place, these data are very difficult to obtain. The data structures have been modeled with the aim of reflecting as much as possible the reality, but because of lack of authentic data, the knowledge obtained from this analysis cannot be very well quantified in terms of value. The great advantage of using a data set that is provided by the service provider (be it masked and anonymized) is that real trends could be identified. However, given the fact that we are talking about data with a high level of security, obtaining it is greatly difficult and for this reason we have opted, as a solution, for a data set generated by GEDIS tool.

3. Using data mining for determining user`s mobility

As presented in article [9], a mining structure constitutes a data source view (commonly combining data from multiple sources) along with associated metadata (such as content type and its distribution), filtered in a manner that serves as a basis for subsequent analysis. Such analysis is conducted by employing mining models that might involve additional filtering (by using distinct sets of input columns or selecting rows that contain specific values only) and aliasing (referencing multiple, individually named copies of the mining structure column set) to facilitate execution of their respective mining algorithms (which, in turn, are utilized to execute predictive queries). Mining algorithms are responsible for identifying patterns and interdependencies in arbitrarily selected data subsets (divided into two partitions – training and testing), yielding forecasts that are used, with certain degree of probability, to extrapolate future events.

To begin with, a view that shows (based on historical data) if that user was mobile while using the telemedicine services (no matter the service used and in which day) was created. We consider a user to be mobile if the distance he has traveled is greater than 0, and in this case, we will assign value 1 to the flag “mobile_cust”. This view will be further used for creating the mining structure and will be considered as source for the mining algorithms applied.

Microsoft SQL Server Integration Services (MSSIS) provides data mining tools for building and browsing through data mining models, by using Business Intelligence Development Studio. The aim of the following SSIS project will be to analyze previous data regarding customer mobility and create some projections based on that.

For the newly created Integration Services project the data source and data source views are set as in figure 1. The table on which we will base the creation of the new mining structure is Customer Details. We can choose as well to have some tables nested, but this applies only in the case of tables for which the relation with the table selected as Case is multi–one relationship. In our case FactServiceUsage would comply. The next step would be to establish what we want to predict, and in this particular case we want to understand if a user is mobile or not, meaning if he is accessing/using telemedicine services while he is traveling distances and not from a fix location. With this purpose, we will select “mobile-cust” as the field we will want to predict and as input for predictions we will select different usage characteristics such as the service used, how accesses were made and how much traffic it generated. We select the columns of data to include in the structure (not all columns need to be added to the model) and define a key (cust_imsi). “Create Testing Set” page serves to specify how much of the data is to be used for training, and how much is to be reserved for use as a test set.

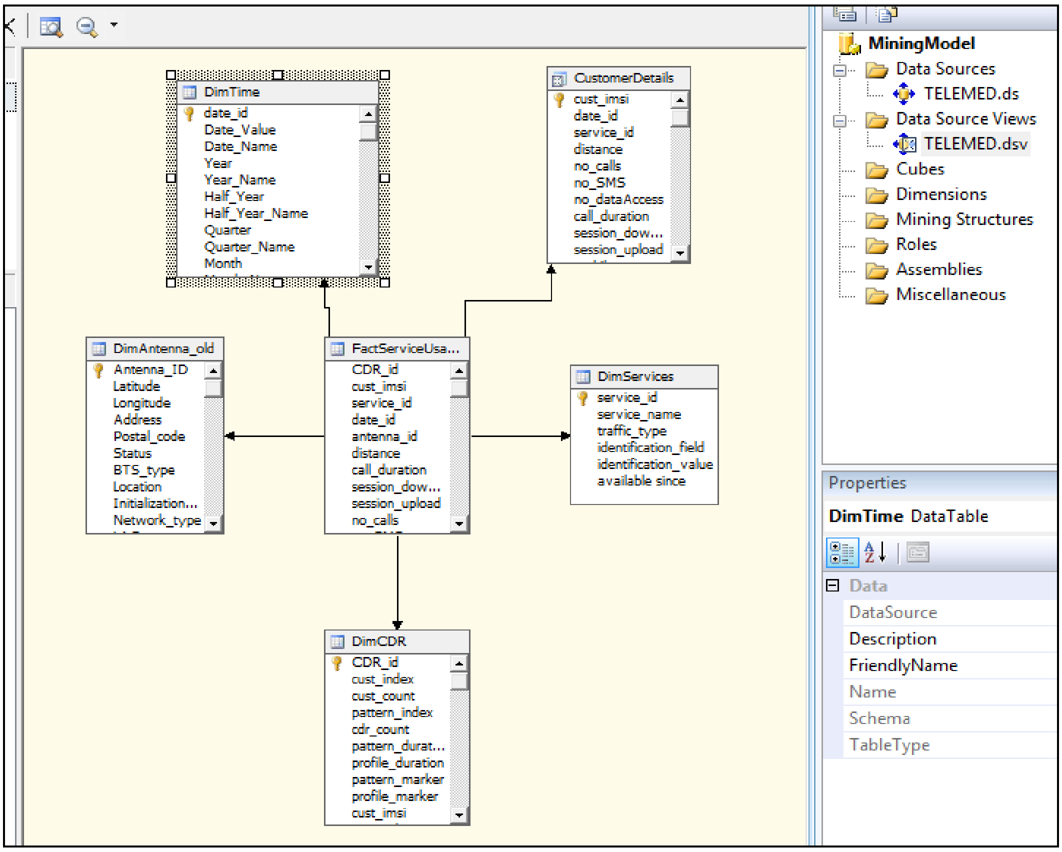
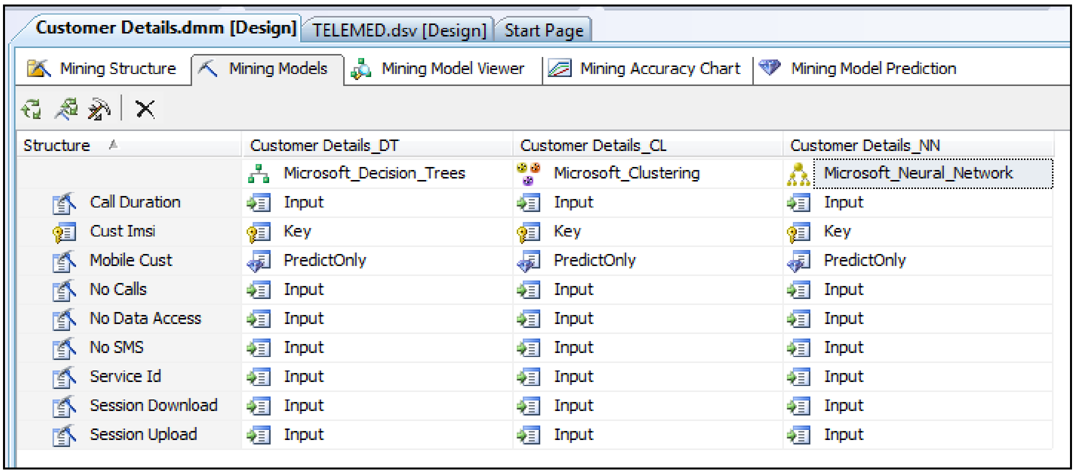


Fig. 1. Setting up the data source views

Separating data into training and testing sets when you create a mining structure makes it much easier to assess the accuracy of mining models that you create later. We will let the default percentage, 30%, of data that we will keep in our back pocket with the purpose of using it for testing the accuracy of the data mining algorithms used. We will see how this will be employed later on in this paper while looking at the “Mining Accuracy Chart”. After processing the mining structure, the data mining models are set. As described in [10], most commonly used techniques in data mining are: artificial neural networks, genetic algorithms, rule induction, nearest neighbor method and memory based reasoning, logistic regression, discriminate analysis and decision trees. Knowledge discovery with data mining is the process of finding previously unknown and potentially interesting patterns and relations in large databases. Future prediction and decision can be made based on the knowledge discovery through data mining.

From figure 2 we can see that “Customer Details_DT” corresponds to the decision tree algorithm. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Other algorithms used are clustering and neural networks (corresponding to “Customer Details_CL” and “Customer Details_NN”). Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. When it comes to neural networks, in practical terms neural networks can be described as non-linear statistical data modeling tools as presented in [11]. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Figure 2 depicts the different mining models that we have selected to use for our mining structure. In order to predict a subscriber`s mobility while using telemedicine services we have selected a three algorithms decision tree, clustering and neural networks.



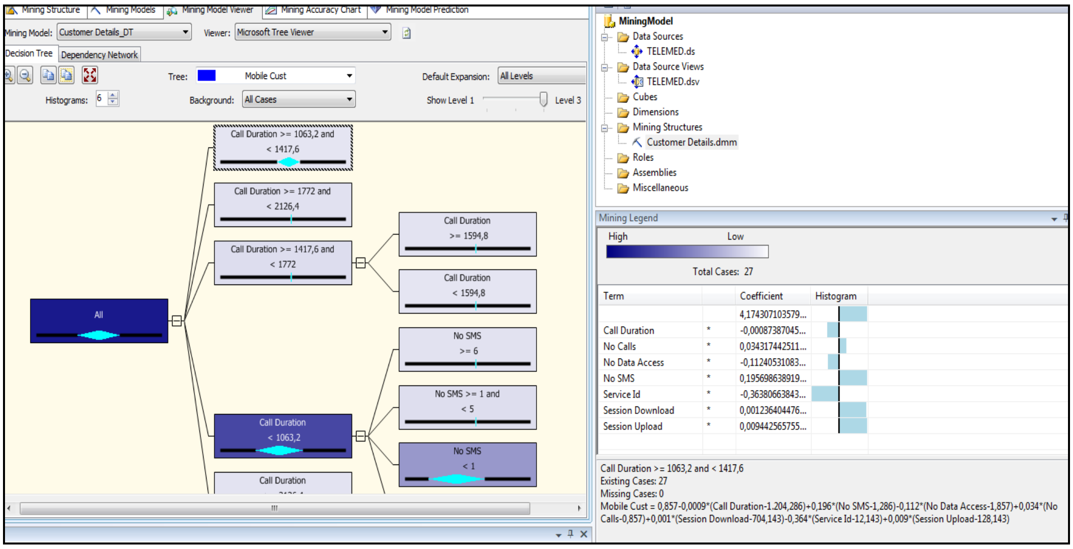
Structure	Customer Details_DT	Customer Details_CL	Customer Details_NN
	Microsoft_Decision_Trees	Microsoft_Clustering	Microsoft_Neural_Network
Call Duration	Input	Input	Input
Cust Imsi	Key	Key	Key
Mobile Cust	PredictOnly	PredictOnly	PredictOnly
No Calls	Input	Input	Input
No Data Access	Input	Input	Input
No SMS	Input	Input	Input
Service Id	Input	Input	Input
Session Download	Input	Input	Input
Session Upload	Input	Input	Input

Fig. 2. Mining models

Exploring the decision tree with the Mining Model Viewer tab we will see that our tree has three levels. For the first level the coefficient is 0.7. If we are drilling down more and look for example on a second level branch (that has call duration bigger than 1063,2 and less than 1417,6) the overall coefficient is 0, 4 and it’s split per each term as it can be seen in figure 3.

Moreover, for the decision tree model we can also visualize which are the nodes used to predict the Mobile Cust node, and the prediction direction (in our case there is no both ways prediction, just one way). The Dependency Network view is an interactive tab that allows eliminating gradually the weakest links.

As a result of doing that, the strongest links in predicting the customer`s mobility are identified to be call duration and the type of service used. If we proceed further in order to determine the strongest link, then we will see that this one is in fact the call duration.



Mining Model: Customer Details_DT Viewer: Microsoft Tree Viewer

Tree: Mobile Cust Default Expansion: All Levels

Histograms: 6 Background: All Cases Show Level 1 Level 3

Total Cases: 27

Term		Coefficient	Histogram
		4,174307103579...	
Call Duration	*	-0,00087387045...	
No Calls	*	0,034317442511...	
No Data Access	*	-0,11240531083...	
No SMS	*	0,195698638919...	
Service Id	*	-0,36380663843...	
Session Download	*	0,001236404476...	
Session Upload	*	0,009442565755...	

Call Duration >= 1063,2 and < 1417,6
Existing Cases: 27
Missing Cases: 0
Mobile Cust = 0,857-0,0009*(Call Duration-1.204,286)+0,196*(No SMS-1,286)-0,112*(No Data Access-1,857)+0,034*(No Calls-0,857)+0,001*(Session Download-704,143)-0,364*(Service Id-12,143)+0,009*(Session Upload-128,143)

Fig. 3. Exploring a branch of the decision tree for predicting customer mobility

So far, we have built and process multiple data mining models based on a single mining structure and now we are ready to create predictions from all related models. As described in [12] , “the Mining Accuracy Chart pane provides tools to help gauge the quality and accuracy of the models you create. The accuracy chart performs predictions against your model and compares the result to data for which you already know the answer. The profit chart performs the same task, but allows you to specify some cost and revenue information to determine the exact point of maximum return. [...] In practice, it is a better to hold some data aside when you train your models, to use for testing. Using the same data for testing that you trained your models with may make the model seem to perform better than it actually does.”

We have in fact kept 30% of our historical data to use it as a test set and we will move further to compare how the three models defined will predict the mobility of a user for that data set. The clustering model seems to have the best performance with a score of 0.72, however being quite far from the ideal model. He is tightly followed by the decision tree model with a score of 0.71.

We can evaluate the clustering model also through the Classification Matrix that shows how many times a model made a correct prediction and what answers were given when the answers were wrong. This can be important in cases where there is a cost associated with each wrong decision. In the case of the clustering model, out of 40 mobile users 49 were predicted to be mobile, while out of 112 non-mobile users 99 were predicted to be non-mobile.

As a result of the mining accuracy chart, there are two options for moving further in obtaining the prediction for customer mobility indicator. Either we can use the clustering model, that had the best results in term of accuracy or we continue searching for another data mining model that could perform better and be closer to the ideal model. For the scope of this study we will use the clustering model and propose to do a search by trying out other data mining models and test them in a future work.

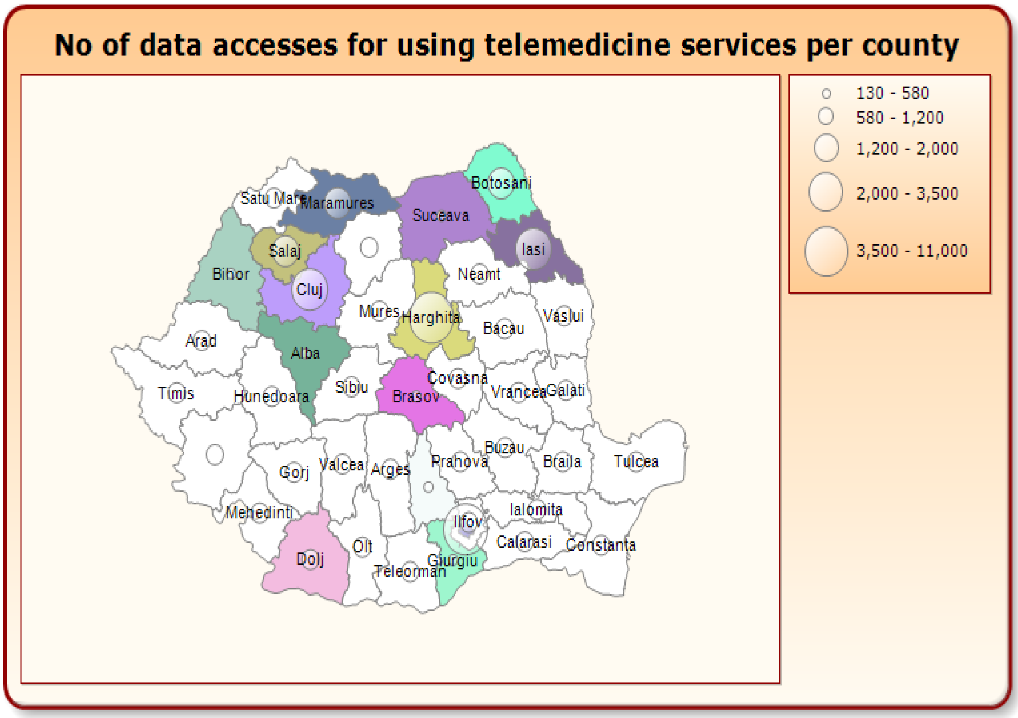


Fig. 4. No of data accesses for using telemedicine services per county

After obtaining this type of indicator, based on the mobility score a customer can be targeted with special designed telemedicine services packages that address his needs.

Moreover, this indicator together with a series of usage indicators (such as number of calls, number of data accesses, intensity etc.) can be easily followed by their geographical distribution) over a chosen period of time. Figure 4 depicts which are the counties with the biggest number of telemedicine services data access and as it can be seen Harghita county is leading followed by Cluj and Iasi. The representation from this dashboard takes in consideration all telemedicine services that have been used over a period of time of five months, from January to May 2015.

The validation of the results obtained by using the clustering model can be done by observing future traffic and the actual user`s mobility while generating that. Comparing the reality with what was predicted, would help in further making adjustments also to the data that enters the study. As this comparison is not possible at the moment, this remains an experimental approach that needs further validations.

4. Conclusion

Analyzing the data generated from using telemedicine services and moreover correlating it with the geographical position where the traffic was made, is a line of research that was not yet pursued and developed. Especially if we are thinking about the context: where most of the studies and researches regarding telemedicine were more focused on improving the delivery of existing services (by using cutting edge technology, new devices, etc.) or developing new services, we can understand that this line of research was not so much exploited.

Exploring telemedicine usage data could be a step forward in improving telemedicine services offering and adding the geographical localization where the traffic was generated is a plus in better understanding the customers and targeting them.

With this purpose, the paper represents a walkthrough the data sources of the study and the glimpse of the methods that could be used for analyzing this data: applying data mining models for predicting different indicators and then representing these indicators as suggestive as possible, making use of spatial analysis, where the case.

Available studies and papers in the area of CDRs and more specifically CDR structure and content are very limited. In an earlier study [4], a consolidated imagine is built based on the existing documentation and a selection of most relevant fields to enter the study is proposed. This paper tackles also a manner in which sample data was generated and how some of the assumptions have been validated by observing the business logic implemented in the GEDIS test data generator.

For the practical part, a solution for analyzing telemedicine usage data is proposed. With this purpose, we have opted for a traditional approach creating a data warehouse, an OLAP cube and several dashboards that query the cube. The technical approach was further described in [4] and [13], while this paper focuses more on one area that was also explored as part of the research: obtaining predictions by using data mining algorithms (subscribers' mobility). In this way we have opened a broader view on the different analysis that can be performed with this solution.

REFERENCE

[1]. *N. Radjou, J. Prabhu, S. Ahuja, K. Roberts*, Jugaad Innovation: Think Frugal, Be Flexible, Generate Breakthrough Growth, Jossey-Bass, 2012.

[2]. *K. M. Zundel*, Telemedicine: History, Applications, and Impact on Librarianship., Bulletin of the Medical Library Association 84, 1996, pp.71-79.

[3]. *T. Molfenter, M. Boyle, D. Holloway, J. Zwick*, Trends in Telemedicine Use, Addiction Treatment., Addiction Science & Clinical Practice, 2015.

[4]. *A. Sarbu*, Optimal Data Architecture for an Telemedicine Analytic Platform., Daaam International Scientific Book 2013, edited by B. Katalinic & Z. Tekic, Vienna, Austria: DAAAM International, 2013, pp.647-654.

[5]. *V.D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, C. Ziemlicki*, Data for Development: The D4d Challenge on Mobile Phone Data., 2013.

[6]. *D. F. Kune, J. Koelndorfer, N. Hopper, Y. Kim*, Location Leaks on the Gsm Air Interface., 19th Annual Network & Distributed System Security Symposium in San Diego, California, 2012.

[7]. *A. Schmidt-Dannert*, Positioning Technologies and Mechanisms for Mobile Devices., Seminar Master Module SNET2,TU-Berlin, 2010.

[8]. *S. Mehta*, Working with Spatial Data Part I – Geocoding Text-Based Spatial Data for Use in Ssis Packages, SQLServerPerformance.com, 2014.

[9]. *M. Policht*, Data Mining Query Task in Sql Server Integration Services, Database journal, 2011.

[10]. *M. Karim, R. M. Rahman*, Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing, Journal of Software Engineering and Applications 6, 2013, pp.196-206.

[11]. *Y. Singh, A. S. Chauhan*, Neural Networks in Data Mining, Journal of Theoretical and Applied Information Technology 5, no. 6, 2009.

[12]. *Z. Tang, J. MacLennan,* Data Mining with Sql Server 2005, Wiley, 2005.

[13]. *A. Sarbu*, Platformă analitică pentru studiul datelor generate de serviciile de telemedicina, Revista Română de Interactiune Om-Calculator (Analytical platform for the study of usage data generated by telemedicine services, Romanian Journal of Human - Computer Interaction - in Romanian), 7 (1), 2014, pp.37-52.

[14]. *P. C. Chapman, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth*, CRISP-DM 1.0 - Step-by-step data mining guide., 2000.