Original Scientific Paper/Originalni naučni ra UDC/UDK: 004.7:004.92]:004.85
Paper Submitted/Rad primljen: 27. 10. 2021.
Paper Accepted/Rad prihvaćen: 21. 1. 2022.
DOI: 10.5937/SJEM2201001J

Novi pristup modelu rudarenja podataka koji je zasnovan na grafovima

Stefana Janićijević[1], Anđelija Mihajlović[2], Anđela Kojanić[3], Aleksa Ćuk[4]

[1]Information Technology School Belgrade, Savski nasip 7, Belgrade, Serbia, stefana.janicijevic@its.edu.rs
[2]Information Technology School Belgrade, Savski nasip 7, Belgrade, Serbia, andjelija37719@its.edu.rs
[3]Faculty of Organisational Science Belgrade, Jove Ilica 7, Belgrade, Serbia, ak20203705@student.fon.bg.ac.rs
[4]Singidunum University, Danijelova 32, Belgrade, Serbia, acuk@singidunum.ac.rs

Apstrakt: Ovaj rad predstavlja novi pristup analize komunikacionih mreža (CNA) koja je interdisciplinarna podoblast naprednog koncepta analize društvenih mreža (SNA). Predlaže se identifikacija relevantnih vrhova unutar povezanih komponenti u grafovima telekomunikacionih mreža . Pored ovog rezultata, algoritam opisuje ponašanje između članova komponente, istraživačke interakcije između komponenti i korišćenje telekomunikacionih usluga . Algoritam se zasniva na kombinaciji dve važne tehnike mašinskog učenja – tehnike klasifikacije Ekstremno povećanje gradijenta (XGB) i algoritma grafa koji se sastoji od Izrezivanja, K-susedstva, Izolovanih ostrva i izračunavanja Mere centralnosti. Ovaj model data mininga se koristi u telekomunikacionim kompanijama kao deo marketinških strategija i procesa upravljanja kampanjama jer se influenseri nagrađuju za doprinos širenju i usvajanju mrežnih usluga među članovima.

Ključne reči: CNA, SNA, Graph algorithm, Pruning, K-Neighbourhood

New Approach for Graph Based Data Mining Model

Abstract: This paper presents new approach of Communication Network Analysis (CNA) that is interdisciplinary subfield of advanced concept of important Social Network Analysis (SNA). Objects in CNA are members of network discovered as vertices that are linked by edges. Identification of relevant vertices within connected components in telecommunication network graphs, such as influencers are proposed. Beside this result, the algorithm describes behaviour between component members, research interactions between components and telecom services usage. Algorithm is based on a combination of two important machine learning techniques - Classification technique Extreme gradient boosting (XGB) and Graph algorithm that consists from Pruning, K-Neighbourhood, Isolated islands and Centrality measure calculation. This data mining model is used in telecommunication companies as part of marketing strategies and campaign management processes since influencers are awarded for contribution in network services spreading and adopting between members.

Key words: CNA, SNA, Graph algorithm, Pruning, K-Neighbourhood

1. Introduction

Graph theory

Graph theory is based on mathematical structures such as graphs. The graph representation is ordered pair $G = (V; E)$, where V is is a finite, non-empty set of vertices (tops, vertices), and E is a set of two-element subsets of set V, i.e. a set of edges (arcs, branches). Graph is a mathematical structure that consists of vertices (vertices or points) which are connected by edges (links or lines). Graph find its application in many areas, from fundamental mathematics, combinatorics, over data science and machine learning.

Graphs could be undirected and directed, depending on whether the edge connecting the vertices u and v is the same as the edge that connects the vertices v and u.

We consider neighbors of every vertex, that is, a set of vertices between which there is an edge. Two vertices are adjacent if there is an edge which connects them, that is, $= \{u, v\} \in E$. For graph G it is denoted set of neighbours that is $N(v) = \{u \in V \mid (u, v) \in E\}$. Degree of vertex v is $d(v)$ and it is the number of neighbours for vertex v. The loop in the graph is the edge which connects the vertex with itself. A graph that has no loop or parallel edge is called prime. Examples of graphs that are investigated in the literature are “zero graph”, ”trivial graph”, ”simple graph”, ”undirected graph”, ”directed graph”, ”complete graph”, ”connected graph”, K-”partite graph”, ”disconnected graph”, ”weighted graph”, ”regular graph”, ”cyclic graph”, ”acyclic graph”, ”star graph”, ”multigraph”, ”planar graph”, etc. (Diestel R 2000).

There are different variants of representing graphs on a computer, and each of them depends on the nature of the problem being solved and the computer resources at its disposal. Under the term computer resources, it mainly refers to the available memory space. Usually, the vertices of a graph are enumerated with $0, 1, 2, \ldots, n-1$ or $1, 2, \ldots, n$, where n is a number of vertices in the graph. The set of edges is represented by one in two ways:

1. The adjacency matrix is represented by the elements that provide information about whether there are edges in between vertices corresponding to the indices of these elements. If the graph is not weighted then the elements of the matrix neighbourhoods only contain information about the existence of an edge between the corresponding vertices. If it is a weighted graph, then the matrix element contains information about the weight of the edges. Formally, the neighbourhood matrix of dimension nxn of G is written as (Janicijevic, 2016):

$$A_G = \left(a_{\{ij\}}\right)_{(i,j)\in VxV} \qquad (1)$$

   Example of the adjacency matrix is:

$$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}$$

2. A graph can be represented by a list of neighbours such that each of the vertices of the graph is an element to which a list is formed in which the neighbours of that vertex are placed in the graph. When working with the sparse graph (a graph in which the number of edges is proportional to the number of vertices), neighbour lists are used because all edges of the graph are described with relatively few data.

However, the list variant is expensive because, for example, a very simple operation such as testing if some two vertices are neighbours requires reviewing the list of neighbours. In practice, the number of such tests can be very large and the performance of the program will be significantly compromised. If they have enough memory, then it is best to present the graph in both ways, to be after they used information either from the neighbourhood matrix or from a list of neighbours, depending on where they were located information that is relevant. The neighbour list is usually made on the basis of distance, so that it is first in lists the vertex neighbour that is closest, and the last vertex the neighbour that is farthest (Amer, 2015).

Well known problems that are solved by graph algorithms are Hamilton loop problem, Graph k-colouring problem, Minimal partition problem, Minimum dominating set problem, Independent set problem, Maximal clique problem, Longest path problem, etc.

Graph applications are wide: Internet, telecommunication networks, scheduling, road network, railway network, power transmission system, etc. Certain problems that are successfully solved are optimal transport of goods, optimal energy transfer, protection against accidents, etc.

Social Network Analysis

Social network analysis (SNA) is also a real-life graph application that finds its purpose in Internet and telecom companies. One of the first difficulties is that this problem belongs to the class of NP-complete (NP-difficult problems), since finding an optimal solution for such problem is almost impossible in real

life time. Solving NP difficult problem with exact methods is expensive and often impossible to do, so it is usually found some heuristic with the help of which an approximate solution is obtained. That solution may not be optimal, but if the heuristics are good enough the solution will be close to optimal. This problem belongs to the problems of combinatorial optimization. For the most part it is possible to write them in the spirit of mathematical programming.

SNA has arisen from the broader field of graph theory and network systems. Interaction between vertices or objects provides approaches for distinguished algorithms to develop insights from communities to predict behaviour of a network.

Social network analysis is analysis of connected objects in any sort of networks and graphs. The Internet is growing and according to that, social networks are growing through the world. We could say that social networks can be viewed as a set of connected entities. Most of the time we represent it by a collection of vertices and edges. A vertex is an abstraction for a user in the network whereas edges are relations between these users (Ahmed and Ismail 2020).

SNA applications provide relations entities and people over the globe. Most social networks are based on friendship and family linkage (Facebook, Instagram, Twitter), but there are also professionally networking (LinkedIn) or advanced expert networks (DNA App, Trading View).

Communication Network Analysis

Large-scale is specialty of telecommunication networks. It is very common to create network with over a million vertices and many more edges. Also, in industry-based cases, number of edges is smaller than the maximal number of edges. A common characteristic of telecommunication graphs is that they are scale free, which means that the distribution of the number of relations of every user follows power law behaviour. One of main characteristics of a network is big data approach. It is very usual to explore networks with millions and millions of users and billions of edges. Another characteristic is sparseness, since complete subgraphs or maximal cliques are very hard to be found for a significant number of vertices (Kihl, 2010).

The process of defining usage communications based on both the network and the graph theory is known as Communication network analysis. Communication Network Analysis (CNA) is the subfield of Social Network Analysis where vertices and edges are common graph-based approach features such as users and relations between them, but conceptually CNA is different than SNA since, edges are developed according to telecommunication traffic such as voice and SMS system.

CNA and SNA methods have become a powerful tool for studying networks in various issues of telco field.

Big Data

Big data occurs in science, industry, and commercial applications. These include government, military, telecommunications, medical, biotechnology, astrology, ecological, pharmaceutical systems, as well as many others. Big data sets are faced with challenges such as data storage, data warehousing, data compression, visualization, information insights, clustering, pattern recognition, algorithm performing. Addressing these issues requires special interdisciplinary efforts in developing sustainable techniques. Big databases imply with them complex and content issues and thus represent a challenging field to address. In many cases, big data sets are represented as large graphs with remarkable attributes that are connected with vertices and edges. Remarkable attributes may contain special information that characterizes the information. Analyzing the structure of such a graph is important to understand the structural characteristics of the application which is represented, such as improving information retrieval and memory organization. Current trends in analyzing large graphs are developed mostly on market graphs, telecommunication graphs and Internet graphs (Ahmed and Ismail, 2020).

2. Problem Formulation

CNA objective is identification of influencers in a network based on the number of connected users and based on the traffic between them, so this paper explores the identification of influencers in the telecommunications network. Regarding to this, there is methodology for identification of the

influencers proposed. This research is based on CNA methodology. The main pillars of CAN are voice, SMS and data traffic.

Msisdn is the mobile station international subscriber directory number and the CAN model is based on it. CNA adopts a record of data produced by the telephone exchange. These data record emanates the details of each Telecom transaction (VOICE, SMS, MMS, GPRS, internet using...) that passes through mobile devices. In conjuction with msisdn, CNA emanates various data sources such as customer data to analyse users' relationships. Coupling this information equable with CAN, "it provides better insights and values that affect the revenue and the customer satisfaction" (Amer, 2015). Msisdn and usage data collected for 4 months. It was implemented ETL (Extract, Transform and Load) to data.

"It is used CNA to analyse relationships among interacting vertices (users, products ...etc.), therefore we discovered the structure of individuals or organizations. Relationships in a network can be directional or non-directional. When we talk about directional relationship, we could say that one person is the initiator (or (source) basis of the relationship) while the other is the one who receives (receiver) (or destination of the relationship)" (Amer, 2015). Weight is indication of the vitality of the connections that can be added.

In CNA most important is the concept that present communications over network flow, such as voice and messaging system. Networks are used to represent group of users or vertices of a network with their linkage characteristics or edges.

3. Related Work

In the paper written by Molhem et al. (2019), the authors are trying to explain what exactly SNA analysis in Telecom data is. "Networks and SNA concepts were applied using Telecom data such as call detail records and customers' data in order to construct a weighted graph in which each relation carries a different weight, representing how close two users are to each other. SNA is used to explore the Telecom network and calculate the centrality measures. Centrality measures help to determine the vertex importance in the network".

"Finding Multi – SIM users within the same operator or across different operators presents another important concern to Telecom companies because it allows improving campaigns and churning models. The paper is based on a real dataset of 3 months MSISDN and customer data provided by a local Telecom operator. Accuracy of 85% was achieved for users from different operators and 92% for users from the same operator".

In the paper written by Pham et al. 2015, the authors survey "recent advances in the study of influencer identifications develop from big data perspectives, and present state-of-the-art solutions of vertices whose removal would breakdown the network. They proposed survey methods to locate the essential vertices that are capable of shaping global dynamics with either continuous or discontinuous phase transitions. The solution implies recommender system in social networks".

In the paper written by Bethu et al 2018, the authors argue that "data science is a concept to unify statistic, data analysis and their related methods in order to "understand and analyze actual phenomena" with data. The principal idea in designing different marketing strategies is to identify the influencers in the network's communication. Targeting influencers usually leads to a vast spread of the centrality measures to identify and assign an influence score to each other. Higher score – better influencer".

The aim is to find the best influencer between users among given pair of users. Algorithm is basically developed to scale all cases over the graph. Data is used for researching pattern and scoring the new data with the prediction for every user among the given pair of users.

3. Model building and validation

Methodology

CNA is a model that explores a user's telecommunications network through the behaviour of each user individually, based on the number of users with whom a specific user comes into contact, but also based on the intensity of communications that user has in the network. The model highlights influencer

users within prominent groups of connected users. In addition, the model provides insights into the mutual relations of network members, but also insights into the relations of users towards telecommunication services.

Main methodology of influencer identification consists from combination of two approaches: Classification algorithm XGB and Graph based algorithm which considers K-Neighborhood, Pruning, Isolated islands and Centrality measures calculation.

Together, they are forming machine learning system for CNA. One of the goals involves modelling the mechanisms that underline human learning. It was developed learning algorithms that are generally consistent with knowledge of the human cognitive architecture and that are also designed to explain specific observed learning behaviors. This machine learning can transform training data into knowledge using algorithm.

The phases of the CNA model are:

- Data preparation,
- Model development,
- Model evaluation.

Data preparation

Data preparation is a very demanding and important process. It implies scrubbing, wrangling, munging and auditing which is performed over tables for A number and for AB numbers communication. These two tables are final and they contain independent and dependent variables that enter the model, which means, selection of variables, deleting redundant variables, working with missing values, editing outliers, normalization of data, etc. Table A number consist from the predictors such as voice count, SMS count, voice duration, gprs count, long voice duration, short voice duration, etc., while the target variable is y.

It is defined according to formula:

$$y = \{1,\ where\ minimum\ 6\ relevant\ predictors\ are\ achieved\ for\ at\ least\ 50\%\ of\ value\ \ 0,\ otherwise\ \} \quad (2)$$

Where:

$$x_i \in X, X \in \{voi_{cnt}, voi_{dur}, sms_{cnt}, data_{cnt}, voishort_{cnt}, voilong_{cnt}, voishort_{dur}, voilong_{dur}, voiin_{cnt}, smsin_{cnt}, voiout_{cnt}, smsout_{cnt}\} \quad (3)$$

Target variable is used for calculation potentially most valuable vertices - influencer vertices with y = 1, so we positively labelled every A number that has more communication than variable median for at least 6 variables predictors.

Model development - Extreme gradient boosting (XGB)

The core of extreme gradient boosting (XGB) itself is the “group algorithm based on the gradient boosting tree. Gradient boosting is an algorithm of boosting in the ensemble algorithm. XGB algorithm is an efficient implementation version of gradient boosting algorithm. Because of its excellent efficiency in application practice, it is a widely-praised technique in industry. XGB is similar to gradient boosting decision tree (GBDT) and is based on the classification and regression tree theory. It is able to build multiple weak evaluators on the data and then summarizes the modelling results of the weak evaluators. In parallel, the XGB model can effectively deal with regression and classification problems to obtain better performance than a single one” (Bhattacharya, 2020).

“Gradient boosting involves the creation and addition of decision trees sequentially, each attempting to correct the mistakes of the learners that came before it. Most implementations of gradient boosting are configured by default with a relatively small number of trees; it is because adding more trees beyond a limit does not improve the performance of the model. The reason is in the way that the boosted tree model is constructed, sequentially where each new tree attempts to model and correct for the errors made by the sequence of previous trees”.

The objective function is formed of two levels. The first level is used to measure the discrepancy between the predicted cost and the actual cost (represents the deviation of the model), and the other part is the regularization term (the variance of the control model). The prediction accuracy of the model is regulated by the deviation and variance of the model.

XGB aims to target and predict a possible influencer. It is used to reduce the user base to a size that is optimal for management, transformation, and manipulation.

1. Initialize $f_0(x) = arg\ arg\ min_\gamma \sum_{i=1}^{N} \quad L(y_i, \gamma)$.
2. For $m = 1\ to\ M$:
   a. For $i = 1, 2, \dots, N$ compute
   $$r_{im} = -\left[\frac{\delta L(y_i, f(x_i))}{\delta f(x_i)}\right]_{f=f_{m-1}}$$
   b. Fit a regression tree to the targets r_{im} giving terminal regions
   $$R_{jm}, j = 1,2, \dots, J_m.$$
   c. For $j = 1,2, \dots, J_m$ compute
   $$\gamma_{jm} = arg\ arg \min_{\gamma} \sum_{x_i \in R_{jm}} \quad L(y_i, f_{m-1}(x_i) + \gamma).$$
   d. Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \quad \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

The goal of XGB is to eliminate vertices with low values of total traffic, and thus to reduce the dimension of the initial graph. Based on the target variable, XGB creates a decision about which user is an influencer.

The final set of users is selected in a new table which is the basis for creating a graph.

XGB is designed for speed and performance. Two reasons to use XGB are execution speed and model achievement. In addition, XGB has proven to be great combination of software and hardware optimization techniques to achieve superior results using fewer computing resources in the shortest amount of time.

XGB consists of following steps:

- Parallelism in tree construction,
- Pruning using the DFS algorithm,
- Computing in external memory,
- Regularization due to reduced overfitting,
- Efficient handling of missing data,
- Built-in cross validation.

Model evaluation parameters:

- Coincidence matrices,
- Performance evaluation,
- Evaluation metric (AUC&Gini).

Model development - Graph algorithm

Graphs are a structure and a powerful tool for modelling and analysing data such as telco and social networks, websites and links, as well as vehicle locations and routes. When there is a set of objects that are interconnected, then they can be represented by graphs.

The graph algorithm aims to observe the interrelationships between influencers and other users on the basis of which the final set of influencers is determined. A graph algorithm is used to highlight measures of significance of each vertex in interaction with other vertices.

The vertices with their edges that formed graph are users that the XGB model has classified as potentially influencers with their connections.

Based on the created table, the weight of the edges between A and B numbers is calculated, defining the calibration parameter for the call length, the number of calls and the number of SMS communications. The greatest weight is assigned to the length of the call, and the least to the number of SMS in accordance with the distribution of the database

$$w = (alpha * 6/8 * row['VOI\_DUR'] + alpha * 2/8 * row['SMS\_CNT'] + (1 - alpha) * row['VOI\_CNT']) \quad (4)$$
$$where\ alpha = 0.5$$

K-Neighborhood

The K-neighborhood method determines the vertex degree, indegree, and outdegree of the 1st level neighbours. They are calculated for each vertex based on this initial graph.

Pruning

"The problem of determining the proper size of an artificial neural network is recognized to be crucial, especially for its practical implementation in such important issues as learning and generalization. One popular approach for tackling this problem is commonly known as pruning and it consists of training a larger than necessary network and then removing unnecessary weights/vertices. The algorithm also provides a simple criterion for choosing the units to be removed, which has proved to work well in practice. The results obtained over various test problems demonstrate that effectiveness of the proposed approach" (Zhang and Mingyang, 2019).

Data set consists from vertex-level information for all vertices that are still included in the graph after removal of weaker edges. Weak edge considers all edges smaller than median weights of all edges.

Isolated islands

Isolated island gives an answer to the standard question: "Counting the number of connected components in an undetected graph". A connected component of an undirected graph is a subgraph in which every two vertices are connected to each other by a path, and which is connected to no other vertices outside the subgraph. A group of connected for example voice calls forms an island (Hanneman, 2005).

The graph is reduced to components, eliminating weak components with the remaining weaker vertices. Using the Isolated Islands method, the graph is divided into mutually isolated components of optimal size while less isolated components are discarded. The graph is divided into connected components and further separation is performed, so that no subgraph contains less than the number of vertices defined through minimum list length. The Isolated islands method separates the components at the level of bond strength, so that the edges with the least weight are eliminated first. Within each stronger component, the central vertices (i.e. the carriers of communications in the group) are isolated. BFS algorithm reduce graph on strongly related components. In BFS, we start from a specific vertex and explore as far along each level of vertices as possible before re-searching backwards. We also need to monitor visited vertices. When we implement BFS, we use the stack data structure to support reverse lookup.

Centrality measures

Metrics are the final part of the model that decides who is the influencer among the vertices. Metrics are calculated for each vertex, and, we could say the metrics calculated the importance of each vertex. Typical metrics that are used the closeness and the betweenness. They are based on shortest paths and distances between vertices (Zhang, 2017).

$$CLOSENESS = \sum_{y \neq x} \frac{1}{d(y,x)}. \quad (5)$$

$$`BETWEENESS = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (6)$$

Targeting influencers usually leads to a vast spread of the centrality measures to identify and assign an influence score to each other.

Higher score means better influencer.

The most important vertices are selected by sorting and comparing the highest values of these metrics, based on previously isolated components for each vertex.

The selection of influencers is made on the basis of the Table 1. These are the strongest users in the network, according to degree and who have many outgoing communications and many incoming communications. In production, the model is realized by hundreds of large groups of connected users, which are shown here.

Table 1. Table of final result after isolated islands and centrality measures calculations

	group id	msisdn	degree	closeness	betweenness
604	301	38*******63	1930	0.002142531	1861383.467
521	902	38*******27	1628	0.002141762	1324349.875
868	87	38*******30	1415	0.002141815	939444.9204
58	149	38*******24	1311	0.002142061	856551.797
871	406	38*******10	849	0.00214199	342381.3873
59	640	38*******00	767	0.002141245	293623.2

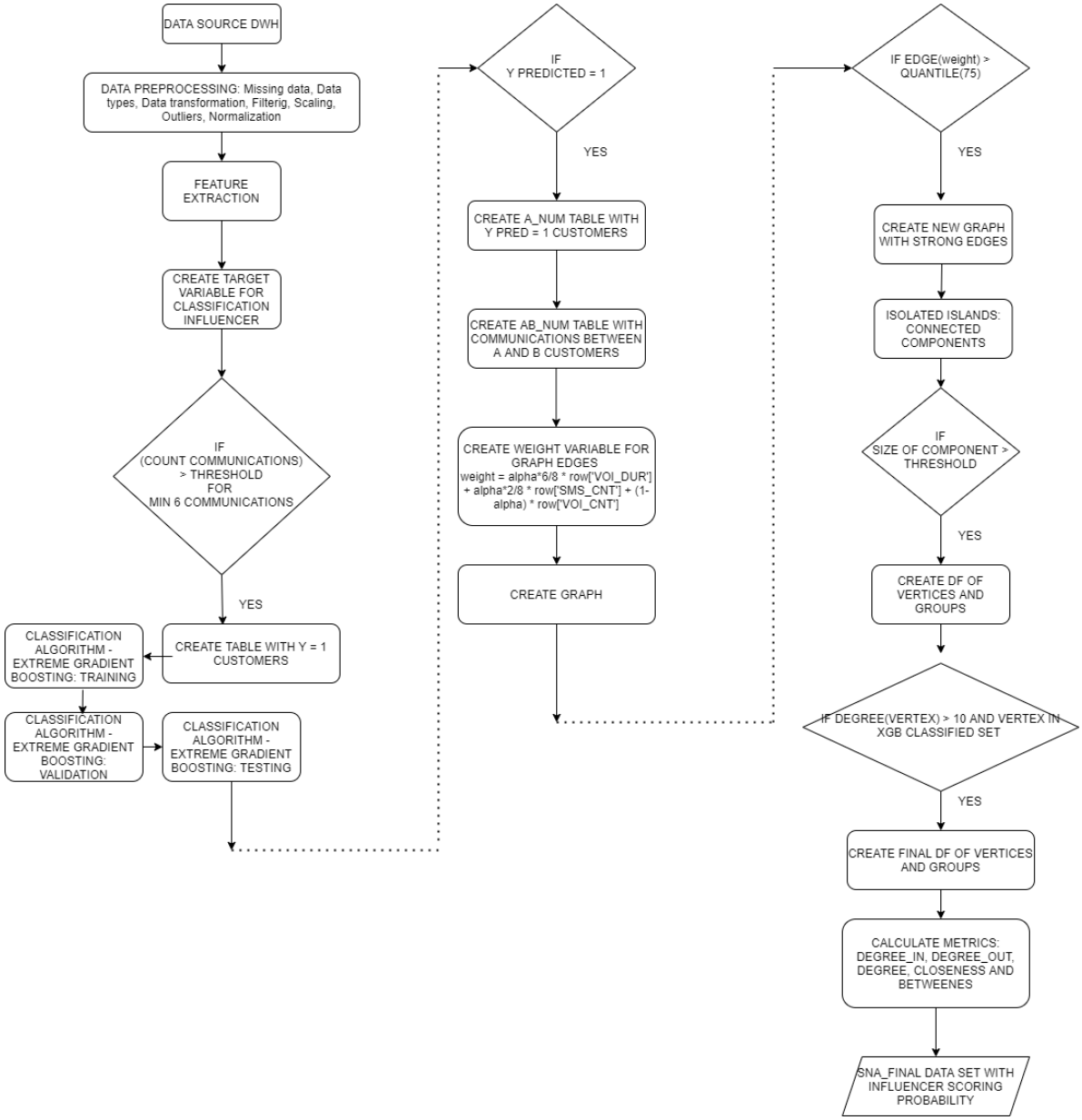

Figure 1. Algorithm CNA Source: Authors

Pseudocode

1. $Initialize\ f_M(x), where\ \ f_M \in f^N\ space\ of\ classified\ subscribers$
2. Create weighted graph and adjacency matrix

$for\ i = 1\ to\ M$
$for\ j = 1\ to\ M$
$x[i,j] = w(i,j)$
$iterate\ through\ all\ connections$

3. $Calculate\ K - Neigborhood$

$for\ i\ =\ 1\ to\ M$
$for\ j\ =\ 1\ to\ M$
$if\ (G{-}{>}\ x[i][j]\ ==\ 1)$
$degree++$

4. $Pruning$

$Input$: $A\ weighted\ graph\ G\ =\ (V,E)$
Output: Subgraph $H \subset G$
1: Sort edges E by weights in an ascending order.
2: $F \leftarrow E$
3: $n \leftarrow (|E| - (|V| - 1))$

4: { Iteratively prune the weakest edge which does not cut the graph }
5: i ← 1, j ← 1 { j is an index to the sorted list of edges }
6: while i ≤ n do
7: if C(u, v; F \ {ej}) is not − ∞ then
8: F ← F \ {ej}
9: i ← i + 1
10: j ← j + 1
11: Return H = (V,F)

5. $Isolated\ islands$

```
for i = 1 to M
    for j = 1 to N
      if (x[i][j] && ! visited[i][j]) {
        // visited yet, then new island found
        // Visit all cells in this island
        BFS(x, i, j, visited)
        // and increment island count
        ++count
```

6. $Calculate\ centrality\ measures$

4. Results and interpretation

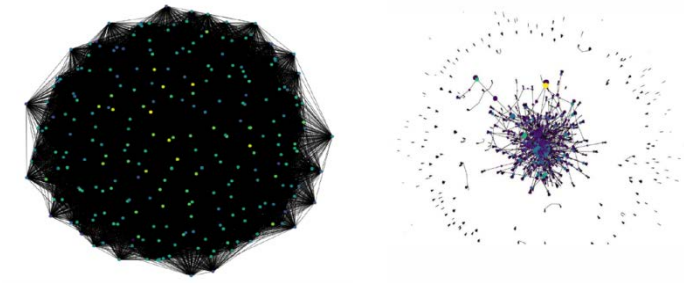

Figure 2. These are population graph and sample graph
Source: Authors

Classification method is trained on 1130925 unique MSISDN data set. Target variable y is equal 0 on 601350 MSISDN and 1 on 529575 MSISDN. After classification, it was calculated contingency matrix distributed on TP, FP, TN and FN as follows:

Table 2. Extreme Gradient Boosting Y - Train set

XGB - Y/TRAIN	**0**	**1**
0	537488	63862
1	58725	470850

Table 3. XGB results for different parameters combinations on train data set

Classification	***AUC***	***GINI***	***Accuracy***	***Precision***
XGB (max depth = 10, boost = 5, min child weight = 0.1, max delta step = 0.7)	0.698	0.720	63.3%	74.83%
XGB (max depth =20, boost = 10, min child weight = 1.0, max delta step = 0.5)	0.713	0.731	65.7%	76.24%
XGB (max depth =25, boost = 10, min child weight = 1.0, max delta step = 0.6)	0.762	0.791	69.4%	79.14%
XGB (max depth =35, boost = 5, min child weight = 1.1, max delta step = 0.4)	0.	0.837	74.3%	82.51%
XGB (max depth =40, boost = 10, min child weight = 1.1, max delta step = 0.2)	**0.904**	**0.802**	**83.16%**	**88.91%**

Best approach is last combination of parameters for this model. Parameters are:

- Tree method: basic,
- Number of boosts: 10,
- Max depth: 40,
- Min child weight: 1.1,
- Max delta step: 0.2
- Objective function: binary logistic,
- Sub sample: 1.0,
- Eta: 0.3,
- Gamma: 0.0,
- Alpha: 0.0
- Scale pos weight: 1.0.

After selection of best combination of parameters on training data, it is applied model on evaluation set. Number of unique MSISDN was 2038410. Comparation between total population and $y = 1$ population according to features distribution:

Table 4. Extreme Gradient Boosting Y - Eval set

XGB - Y/EVAL	**0**	**1**
0	971640	132346
1	111081	823343

Table 5. Comparing $\$XGB$ with y

Correct	1,794,983	88.06%
Wrong	243,427	11.94%
Total	2,038,410	

Table 6. Evaluation Metrics

Model	**AUC**	**Gini**
%XGB - y	0.907	0.815

For the evaluation set it is achieved 16234 strong vertices (influencers) in 2341 isolated components.

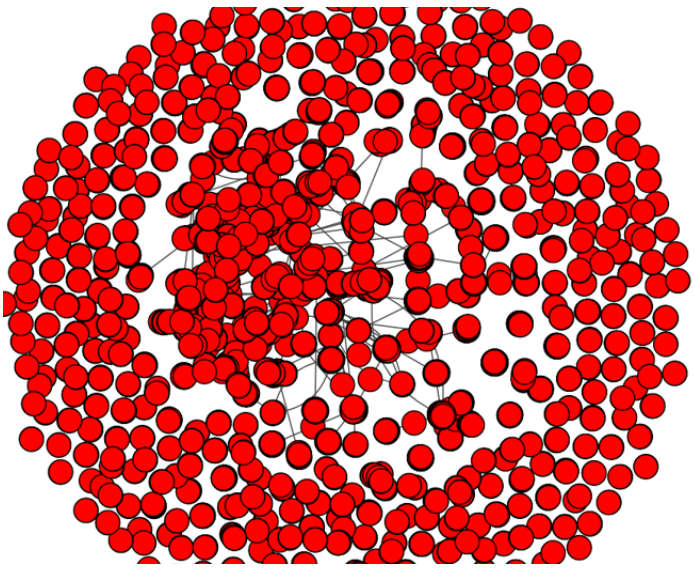

Figure 3. This is influencer graph
Source: Authors

Conclusion

Global graph search is currently the most optimal method of processing large data; all large systems such as the Internet and telecommunications use graph-based algorithms since other heuristics are almost impossible without it. The CNA model provides and regularly submits a list of users and groups sorted by the highest probability. A list of users from the most influential to the least influential according to K-Neighbourhood degree in the selected group of influencers is submitted. For related groups, information on group size and strength is provided, which is also sorted by probability from largest to smallest.

The results of the model present possibilities for telco industry and for advanced analytics.

The model approach is very well applied for the big data concept for HDFS but also for DWH concept. It is possible to research telco sparse graphs and calculate different thresholds for pruning weak edges and threshold for the number of components. Based on that table, it is possible to form reports at the aggregate level, as well as at the individual level.

The questions that the model answers are:

- Who are the most influential users - central network users?
- Which groups can be spotted in the network?
- In what way is the network divided into less loosely connected groups?
- How is the network evolving?
- Will the network be maintained or cease to exist?
- How do ideas-information spread through the network?

References

1. Al-Molhem, N. R., Rahal, Y., Dakkak, M. (2019) Social network analysis in Telecom data. *J Big Data 6, 99* https://doi.org/10.1186/s40537-019-0264-6
2. Amer, M. S. (2015) Social network analysis framework in Telecom, *Int J Syst Appl Eng Dev 9.1*, 201-205.
3. Bethu, S. et al. (2018) Data science: Identifying influencers in social networks, *Periodicals of Engineering and Natural Sciences 6.1*, 215-228.
4. Bhattacharya, S. S. S. R. K., Maddikunta, P.K.R., Kaluri, R., Singh, .S, Gadekallu, T.R., Alazab, M., Tariq, U. (2020) A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU. *Electronics*. 2020: 9(2):219. https://doi.org/10.3390/electronics9020219
5. Diestel, R. (2000) *Graph Theory*, Second Edition, Springer-Verlag, New York.
6. Hameeza, A., Ali Ismail, M. (2020) Towards a Novel Framework for Automatic Big Data Detection, in *IEEE Access*, vol. 8, pp. 186304-186322, doi: 10.1109/ACCESS.2020.3030562.
7. Hanneman, R. A., Riddle, M. (2005) Introduction to social network methods. Riverside, CA: University of California, Riverside. Retrieved on January31, 2021 from http://faculty.ucr.edu/~hanneman/
8. Janicijevic, S. (2016) *Variable Formulation and Neighbourhood Search Methods for the Maximum Clique Problem in Graph*, Ph.D. thesis, University of Novi Sad, Faculty of Technical Sciences
9. Kihl, M., Ödling, P., Lagerstedt, C., Aurelius, A. (2010) Traffic analysis and characterization of Internet user behavior, *International Congress on Ultra Modern Telecommunications and Control Systems*, 2010, pp. 224-231, doi: 10.1109/ICUMT.2010.5676633.
10. Pham, T.N., Li, X., Cong, G., Zhang, Z. (2015). A general graph-based model for recommendation in event-based social networks. *2015 IEEE 31st International Conference on Data Engineering, 567-578.*
11. Zhang, J. et al. (2017) Degree centrality, betweenness centrality, and closeness centrality in social network, 2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM2017), Atlantis Press.
12. Zhang, M. et al. (2019) Graph pruning for model compression, arXiv preprint arXiv:1911.09817. Retrieved on March 2021 from https://arxiv.org/abs/1811.08589