A Novel Method for Automatic Discovery, Annotation and Interactive Visualization of Prominent Clusters in Mobile Subscriber Datasets

Shabana K M and Jobin Wilson

R&D Department, Flytxt

IEEE Ninth International Conference on Research Challenges in Information Science (RCIS'15)

13 – 15 May 2015, Athens, Greece

flytxt®
DATA >>>> ECONOMIC VALUE

Flytxt Overview – About Us

Vision, Mission & Impact

- Leading provider of Big Data Analytics solutions that enable Communication Service Providers (CSPs) to derive measurable economic value from subscriber data
- Our vision is to create >10% measurable economic value for CSPs through Big Data Analytics
- Flytxt solutions increase revenues, margins and customer experience for CSPs
- Headquartered in Netherlands, Corporate office in Dubai, Global Delivery Centre's at Trivandrum and Mumbai; and presence in London, Kuala lampur, Lagos, Nairobi, Dhaka & New Delhi

Customers (50+ Customers, 32 Countries)

CSPs/SI

vodafone, airtel, MTN, রবি, banglalink, !dea, MASCOM, AIRCEL, etisalat, IBM, HUTCH, warid, Nepal Telecom, glo, vodacom, Lonestar Cell, MTN, visafone

Brands

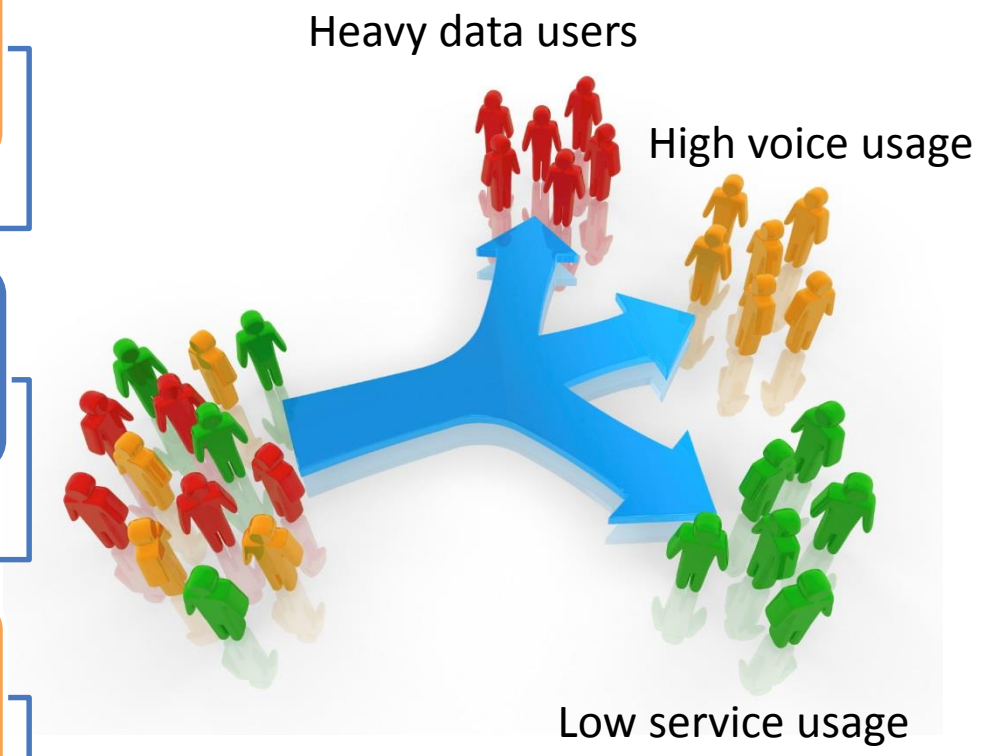
SAMSUNG, P&G, WATCHMAN, marico, Persona, NOKIA Connecting People

Subscriber Segmentation

Process of dividing customers into groups having similar characteristics, behavior or needs

Helps adopt specific marketing plans and promotional offers based on the characteristics and value of each target segment

Improves the overall business performance while enhancing customer experience



Existing Approaches

METHOD	DESCRIPTION	DRAWBACKS
Cluster Analysis	Popular tool that performs automatic discovery of the naturally occurring behavior groups in the data, thereby forming segments that reflect the actual customer behavior trends	a) Number of clusters has to be determined by user b) Requires manual interpretation of final segments
Kandogan[1] [2012]	Cluster annotation performed using metrics such as the density of data values in a cluster, overlap with other clusters, number of outliers in a cluster, and strength of trends that might exist in a cluster	Features have highly skewed distributions in mobile subscriber data sets
Tsiptsis et al.[2] [2011]	Principal component analysis (PCA) performed on the segmentation fields followed by a Two Step clustering Cluster labelling done by identifying correlations of the extracted principal components with the original segmentation fields	Manual effort required to identify associations between the principal components and the original segmentation fields to perform cluster annotation

Motivation

Marketers work on grids where attributes are divided into categories such as Low, Medium, High, etc. Slice and dice performed to arrive at the final segments

Cluster labelling becomes effective when the clusters are consistent or dense with respect to at least one segmentation attribute

This calls for a segmentation scheme that works on category values of attributes and produces clusters that are coherent on at least one attribute

An Outline of our Proposed Approach

In our proposed approach, segmentation is done at two levels:

- Based on subscriber value
- Based on subscriber behavior

Value segments are first created by segmenting subscribers based on a value attribute

The behavior attributes are binned based on user definitions and the absolute values of attributes are replaced with numeric categorical values

Each of the value segments are then clustered using the categorical attributes

The final clusters are automatically annotated based on the most prominent attribute(s) in each cluster

Interactive visualization of the discovered segments is enabled using a tree-map/ doughnut visualization

Our Proposed Approach

I. Subscribers are segmented based on a value attribute selected by the user

- For example, if Average Revenue per User (ARPU) is selected as the value attribute, then the user could define bins like:

BIN Range	BIN Label
0 - 100	Low
100 - 200	Medium
200 - 400	High
>400	Very High

All subscribers with ARPU less than 100 fall in the Low ARPU value segment, those with ARPU between 100 – 200 fall in the Medium ARPU value segment and so on.

We now have divided the subscribers into ***four*** value segments.

Our Proposed Approach (Contd..)

II. Binning Behavior Attributes

- Examples of behavior attributes include Days Since Last Recharge, Outgoing Minutes of Usage (OG-MOU), Average Recharge, etc.
- The attribute values are replaced by categorical values before performing clustering
- The bin values are chosen such that the lowest and highest bin values of all the attributes are the same and the interval between successive bins increases exponentially as the level/value of bin increases
- Bins chosen in this manner were found to produce clusters consistent on at least one attribute, as compared to a binning scheme that used bin values with equal interval
- The lowest bin value is set as 0, followed by powers of 10
- Since the behavior attributes are usually split into a small number of finite bins, the value of the largest bin is not expected to grow beyond a certain limit

Our Proposed Approach (Contd..)

Computation of Bin Values

Let the number of bins for the ith attribute be n_i and the number of behavior attributes be M

1. Compute P

$$P = \prod_{i=1}^{M} (n_{i\_1})$$

2. For each attribute, compute S_i

$S_i = \{ k * (P/ n_{i\_1}) \mid k = 0,1,\ldots(n_{i\_1})\}$

The elements in each Si are sorted in increasing order

3. Compute U

$$U = \bigcup_{i=1}^{M} S_i$$

4. Sort the elements of U in increasing order and compute the mapping

$$M(U_1) = 0$$

$$M(U_i) = 10i\text{-}1 \text{ for all } 2 \leq i \leq |U|$$

5. For each attribute, bin values,

$$B_i = \{M(x) \mid x \in S_i \}$$

where elements of B_i are sorted in increasing order

Consider three attributes divided into 3, 4 and 5 bins respectively (M = 3)

1. P = (3-1)*(4-1)*(5-1) = 24

2. $S_1 = \{ 0, 12, 24 \}$
   $S_2 = \{ 0, 8, 16, 24 \}$
   $S_3 = \{ 0, 6, 12, 18, 24 \}$

3. U = { 0, 6, 8, 12, 16, 18, 24}

4. M(0) = 0 M(6) = 10 $M(8) = 10^2$
   $M(12) = 10^3$ $M(16) = 10^4$ $M(18) = 10^5$
   $M(24) = 10^6$

5. $B_1 = \{ 0, 10^3, 10^6 \}$
   $B_2 = \{ 0, 10^2 , 10^4 , 10^6 \}$
   $B_3 = \{ 0, 10, 10^3, 10^5, 10^6 \}$

Our Proposed Approach (Contd..)

III. Clustering Value Segments

a. Discovery of Initial Centroid Seeds

Discovered through ***frequent pattern mining***

1. For each attribute, except for the bin with the highest value in the segment, all bins are converted to zero.

   For example, under this scheme, (0, 100, 10000,1000, 100) is transformed to (0, 0, 10000, 0, 0), where 10000 corresponds to the highest value bin for all attributes in the segment
2. The unique patterns in the transformed data are mined along with their frequency of occurrence
3. The set of the most frequent patterns whose total number of occurrences is at least a fixed threshold (say 80%) of the total number of patterns in the data set are chosen as the initial centroid seeds

The threshold to select the number of patterns acts as a meta-parameter for the number of clusters, k

b. Clustering of Binned Data

The binned data is clustered using k-means algorithm with Manhattan distance metric. The discovered patterns are used as the initial centroid seeds

Our Proposed Approach (Contd..)

IV. Annotation of Clusters

Cluster annotation is performed by finding the attribute(s) in each cluster which has the maximum uniformity or uniformity above a fixed threshold

1. For each attribute, the bin value with the maximum frequency of occurrence is noted
2. Among all the attribute bins selected in the previous step, the attribute bin with maximum value and the ones with frequency greater than a fixed threshold (say 80%) of the cluster size are recorded
3. The attribute name and bin name of the attribute bins selected in the previous step are combined to form cluster annotation

For example, in a cluster of size 10000 subscribers, if 9000 subscribers have 'High' bin value for SMS and 9750 subscribers have 'Medium' bin value for 'OG-MOU', then with the threshold set as 90%, the cluster would be labelled as 'High SMS and Medium OG-MOU'

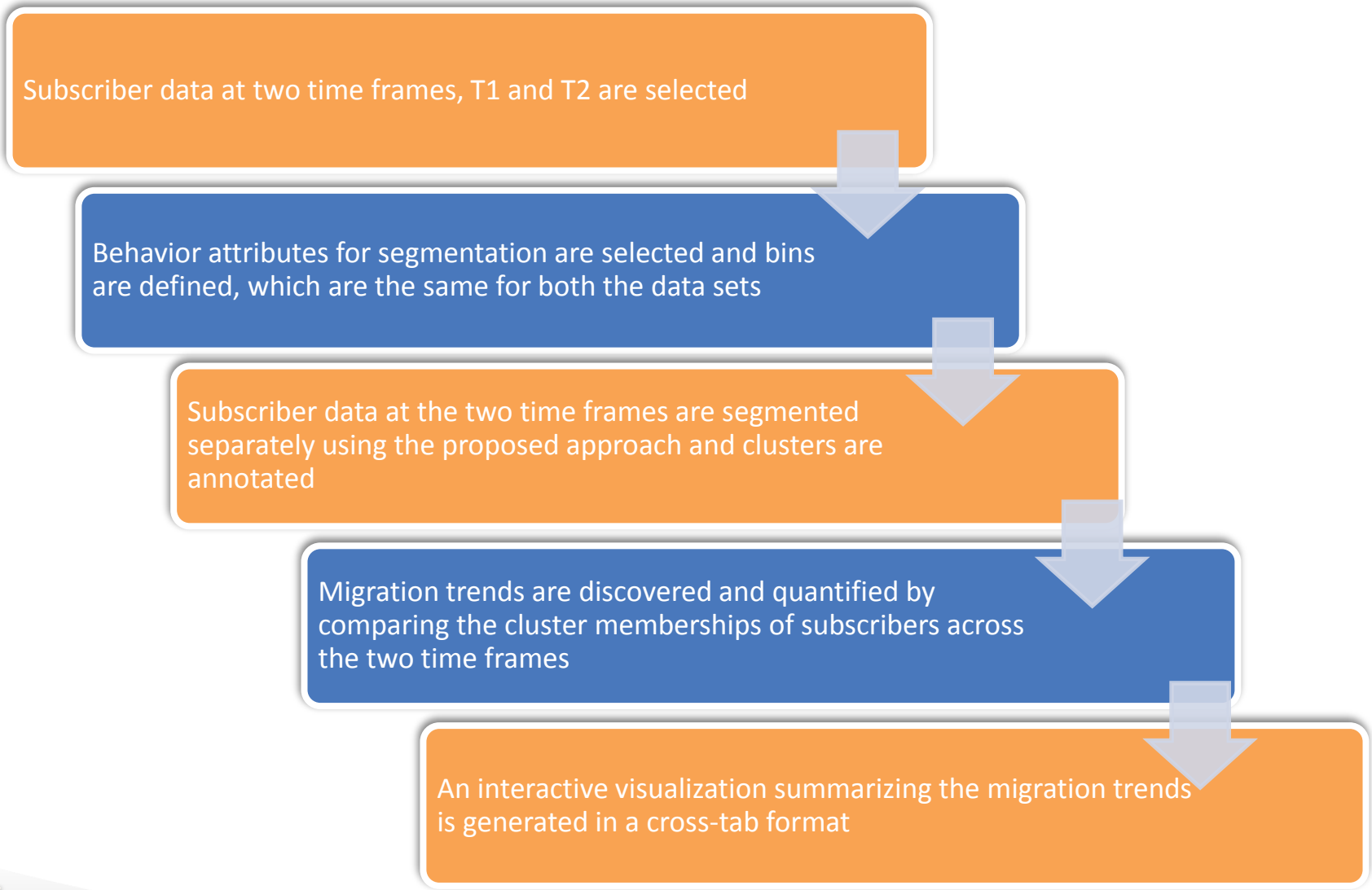
Our Proposed Approach (Contd..)

V. Visualizing Segmentation Results

An interactive visualization in the form of a tree view/doughnut chart is generated

- Summarizes the discovered clusters
- In tree view visualization, clusters are represented as rectangles whose area is proportional to the number of subscribers in the cluster
- Cluster labels are also added to the rectangles
- Rectangles are colored based on the average value of 'value attribute' in the cluster

Segment Migration Analysis using the Proposed Approach



Results

Value Segmentation

The High Value Customer (HVC) segment of a Tier 1 Asian operator's circle consisting of about 0.7 million subscribers was analyzed

- Gross ARPU (Average Revenue per User) was selected as the Value attribute.
- The behavior attributes, also referred to as behavior KPIs (Key Performance Indicators), selected include:

1. Age on Network (AON)
2. Average Recharge
3. Count of Recharge Done
4. Data Usage
5. Out-going Minutes of Usage (OG MOU)
6. On Net Share
7. Voice Out going Call Days

- Three value segments were formed associated with the Low, Medium and High bins of Gross ARPU

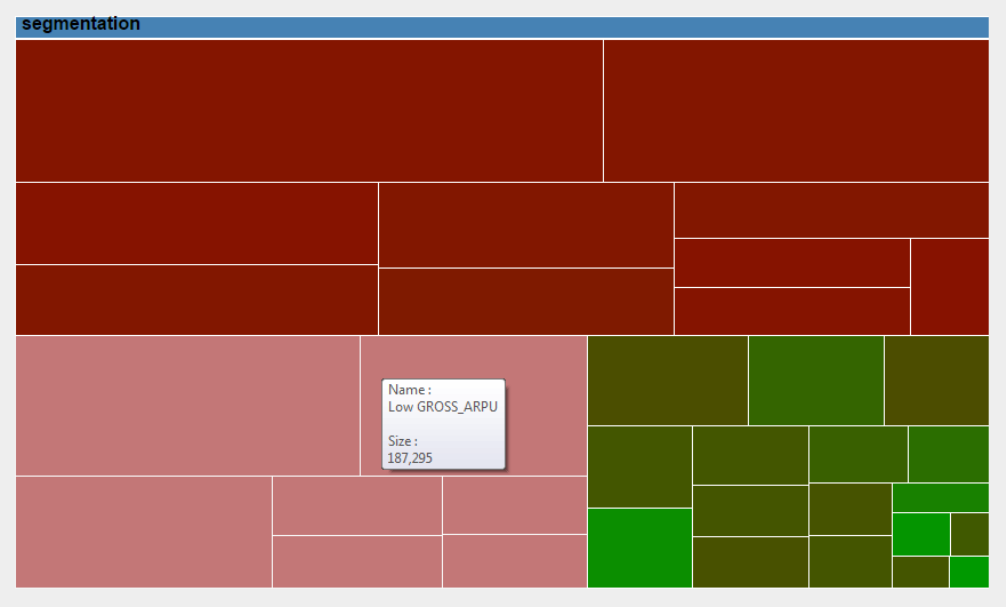
Behavior Clusters in Low Gross ARPU Segment

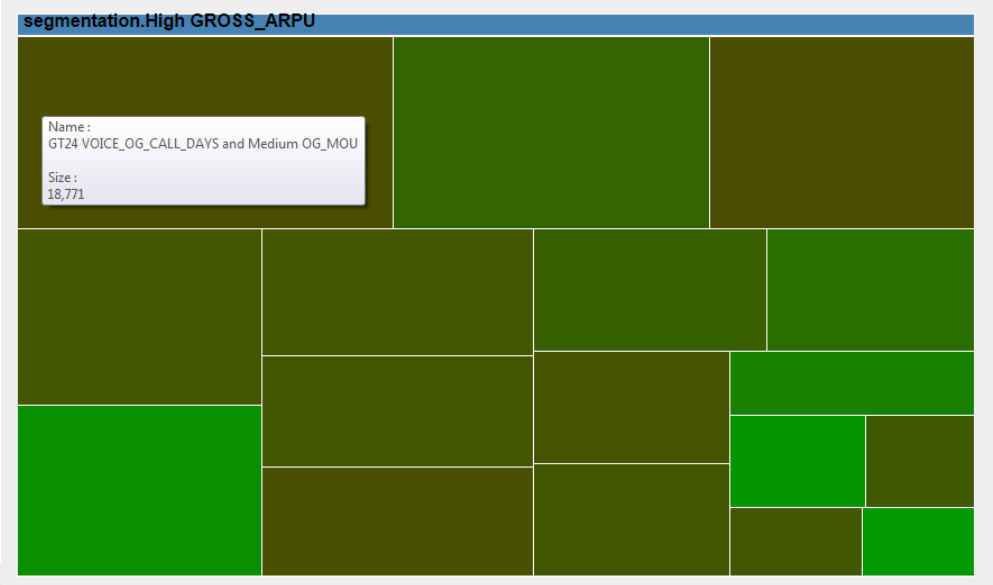
Cluster	Size
GT24 Voice OG Call Days	**63199**
No Data Usage	**37199**
GT2Years AON and ***GT24*** Voice OG Call Days	**41488**
High Data Usage and ***GT24*** Voice OG Call Days	**11353**
GT2Years AON	**13103**
High Data Usage	**10024**
Very High Count of Recharge Done and ***Minimum*** Average Recharge	**10929**

Bin Occupancy of Subscribers in Clusters of Low Gross ARPU Segment

Cluster Label	KPI	0	10	100	1000	10000
High Data Usage and ***GT24*** Voice OG Call Days	AON (Age on Network)	5599	-	5754	-	0
	Average Recharge	7509	3639	-	164	41
	Count of Recharge	4167	-	5965	-	1221
	Data Usage	0	-	0	-	**11353**
	OG-MOU (Outgoing Minutes)	5022	-	6188	-	143
	On net Share	4118	4288	-	2045	902
	Voice OG Call Days	0	-	0	-	**11353**
Very High Count of Recharge Done and ***Minimum*** Average Recharge	AON (Age on Network)	6817	-	4112	-	0
	Average Recharge	**10926**	3	-	0	0
	Count of Recharge	0	-	0	-	**10929**
	Data Usage	9833	-	1096	-	0
	OG-MOU (Outgoing Minutes)	3297	-	7528	-	104
	On net Share	4962	2917	-	1635	1415
	Voice OG Call Days	1559	-	1982	-	7388

Tree map Visualization of Final Segments



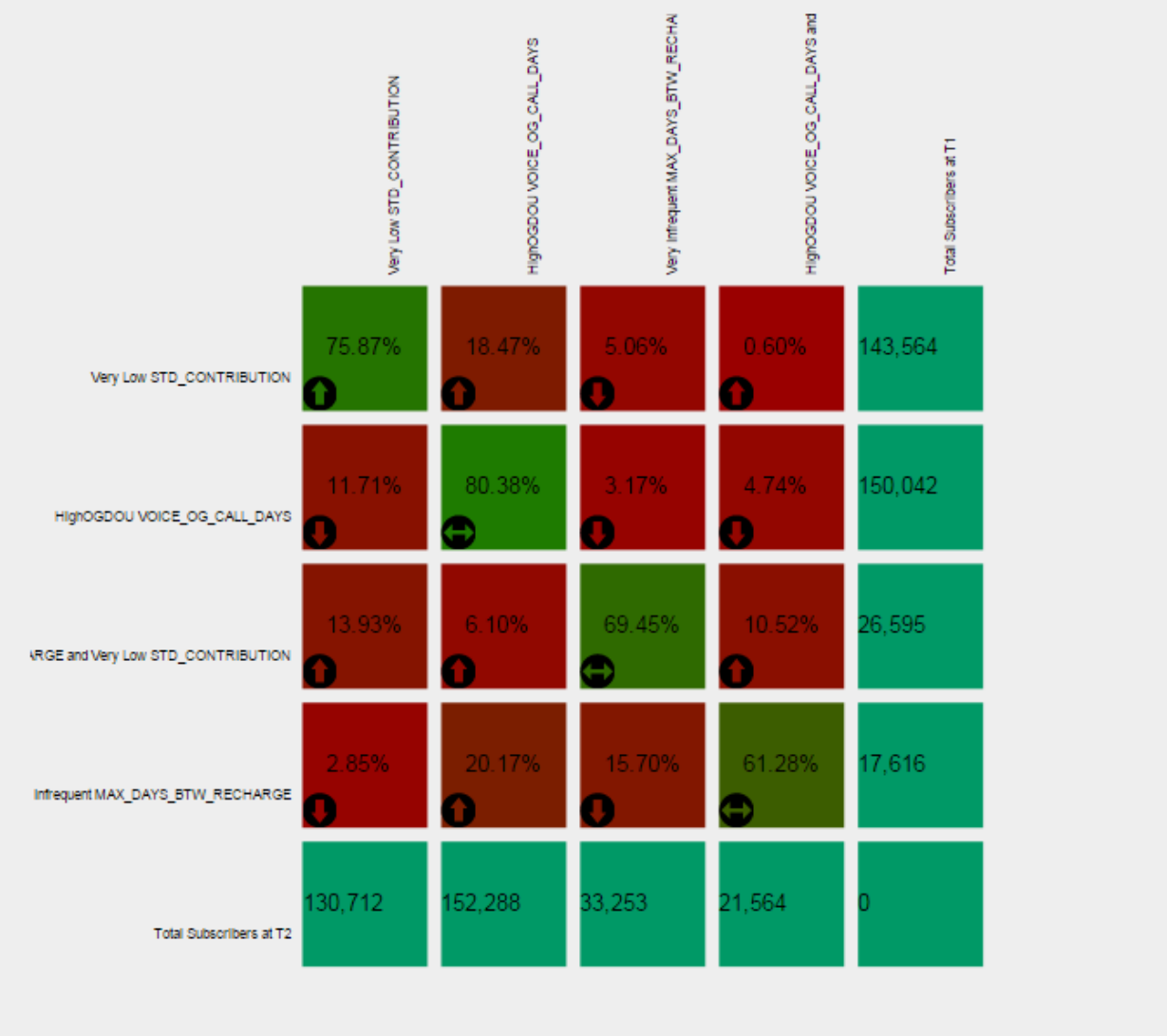


Intra Cluster Cohesion

- In categorical data clustering, cohesion of a cluster is defined as the summation of the cube of the highest probability for each dimension [3]
- Intra cluster cohesion for a clustering result is defined as the average weighted cohesion of each cluster
- The cohesion values ranges between 0 and 1, with values closer to 1 indicating a tighter cohesion

Segment	0.7 million data set	1.8 million data set
Segment -1	0.382	0.899
Segment -2	0.452	0.693
Segment -3	0.569	0.675
Total	0.456	0.781

Segment Migration



Positive and Negative Migratory Trends

Positive Trend

KPI	Average at T1	Average at T2	Difference (T2-T1)
IN DEC	33.85	115.84	**82**
Voice OG Call Days	8.00	21.46	**13.46**
STD Contribution	0.09	0.12	0.03
Maximum Days Between Recharge	12.90	10.38	-2.51
On net Share	0.17	0.23	0.05
OG-MOU	54.50	228.25	**173.75**

Negative Trend

KPI	Average at T1	Average at T2	Difference (T2-T1)
IN DEC	98.28	46.57	**-51.71**
Voice OG Call Days	20.93	10.42	**-10.52**
STD Contribution	0.11	0.10	-0.01
Maximum Days Between Recharge	9.70	11.44	1.73
On net Share	0.22	0.20	-0.02
OG-MOU	174.85	56.91	**-117.95**

Conclusion

- A scalable method for subscriber segmentation has been proposed in which:
  - discretization of attributes is performed before clustering such that the final segments possess good coherence on at least one attribute
  - the number of segments are automatically discovered through frequent pattern mining of the discretized data set
  - final clusters are automatically annotated
  - segmentation results are visualized using tree map/dough nut visualization

- The proposed segmentation scheme could be extended to identify prominent temporal migration patterns of subscribers
- This technique could be applied for customer segmentation in other domains as well
- The areas for future work include formulating a more optimized binning scheme and initial centroid seed selection strategy for clustering

References

[1] E. Kandogan, “Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations,” in Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on. IEEE, 2012, pp. 73–82.

[2] K. Tsiptsis and A. Chorianopoulos, Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons, 2011.

[3] C.-H. Chang and Z.-K. Ding, “Categorical data visualization and clustering using subjective factors,” Data Knowl. Eng., vol. 53, no. 3, pp. 243–262, 2005.

Thank You

www.flytxt.com