Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 144 (2018) 277–290

Procedia Computer Science

www.elsevier.com/locate/procedia

INNS Conference on Big Data and Deep Learning 2018

Customers Segmentation in the Insurance Company (TIC) Dataset

Wafa Qadadeh[a,*], Sherief Abdallah[b]

[a]*The British University in Dubai, Dubai PO Box 345015, United Arab Emirates*
[b]*University of Edinburgh, Edinburgh, UK*

Abstract

Customers' Segmentation is an important concept for designing marketing campaigns to improve businesses and increase revenue. Clustering algorithms can help marketing experts to achieve this goal. The rapid growth of high dimensional databases and data warehouses, such as Customer Relationship Management (CRM), stressed the need for advanced data analytics techniques. In this paper we investigate different data analytics algorithms, specifically K-Means and SOM, using the TIC CRM dataset. While K-Means has shown promising clustering results, SOM has outperformed in the sense of: speed, quality of clustering, and visualization. Also we discuss how both techniques segmentation analysis can be useful in studying customer's interest. The purpose of this paper is to provide a proof of concept (based on a small publicity of data) of how big data analytics can be used in customer segmentation.

Keywords: data mining; data analytics; big data; clustering; CRM; segmentation

1. Introduction

Companies nowadays are continuously working to increase their competitiveness. The availability of big data for Customer Relationship Management (CRM) and data warehouses, with high dimensions, the need to use data mining advanced technologies has been increasing significantly. The usage of data mining algorithm might help businesses to find interesting knowledge in its customer's data both demographic and behavioral then it is the marketing

* Corresponding author. Tel.: +9-7150-792-0868.
E-mail address: wafa.qadadeh@gmail.com

10.1016/j.procs.2018.10.529

experts' responsibility to use these insights in designing the company marketing campaigns to fulfill customer's interests.

The insurance company dataset (TIC), which we mine in this paper, was used in the COIL 2000 challenge. The goal of the challenge was to predict customers who are interested in a caravan insurance policy. The main target of our study is to answer the following research question: Can we discover meaningful clusters using different cluster analysis techniques applied on the high dimensional TIC CRM data? We attempt to answer this question by exploring cluster analysis. We identify some interesting patterns that can be used by marketing experts in insurance companies. In particular, we investigate two different data mining techniques: the infamous K-means clustering algorithm combined with the SOM technique (based on ANN). Using this solution had been showing promising results in clustering CRM dataset and visualization. Clustering and visualization of a high dimensional dataset would be used to recognize the characteristics of a customer in CRM data to design customer centric marketing plans.

The remaining parts of this paper are organized as follows. Section 2 reviews other researches related to using data mining tasks and techniques in CRM data in different domains, showing points of strengths and interests in other high quality pa-pers. Section 3 is the methodology and techniques used in this research. Section 4 proposes our solutions as an experimental evaluation of the used data mining techniques showing the results of applying these techniques on the TIC CRM dataset. Finally Section 5 concludes this study with a general discussion about the proposed solutions and the future work that could be done.

2. Literature Review

Recently, a structured framework was developed to apply Recency, Frequency, and Monetary (RFM), customer's lifetime value (LTV) models [1]. The framework used customers' demographic data to segment banking customers and design marketing strategies. The analysis study consisted of two main phases: In the first phase the CRM data was used to cluster the customers. In the second phase the demographics data variables (age, education, and occupation), which was chosen by SOM technique, was used to re-cluster the resulted segments from step one. Both of these steps had been done using K-Means clustering technique. The customer value comparison used LTV instead of inter/intra cluster distances, in order to maximize the value of the customer, which is one of the targets of this study.

More recent work proposed a Bank Customer segmentation framework, based on customer's LTV [2]. It is very common to study customers segments based on their requirements or preferences. But this study had been handling different approach using the customer's lifetime value which could be more efficient and practical. The researcher had prepared a framework to segment the customers, calculate each segment lifetime value, and estimate the future value of each segment. Two levels of clustering had been implemented on a big dataset of customer's transactions. The transaction record includes deposit type, transaction date, balance before transaction, amount of transaction, etc. K-means and two step clustering algorithms were implemented [2]. The customer's lifetime value had been calculated using a RFM model, the simplest and most powerful customer's LTV approximation model. Finally the study had been using a time series method (multiplicative seasonal ARIMA regression) to predict future value for each segment [2].

In [3] another analysis study was conducted to segment bank customers based on their behaviors to help the bank to prepare retention strategies and gaining new customers. The dataset in [3] is an integration of three tables. First, the customers' demographic data table including age, gender, marital status, and etc. Second, the transaction table containing the customers' transactions. Third, the cards table includes the data for bank cards. Many important information was taken into consideration during this study and their attributes were combined with other customer's attributes such as: transaction type, frequency of a transaction, service type, bank type, and channel type (ATM, Web, and Terminal). The author had classified the factors based on their profitability using ANN.

In [4] investigated the problem of identifying potential customers in big datasets. The study had followed the following methodology: first for accuracy, a semi-supervised techniques were used to build customer behavior modeling automatically. Second, the authors had used a neural network technique to visualize data. For the semi-supervised proposed technique, a multi-layered perceptron neural network with back propagation was used. The classifier is re-trained using subset of labeled data each time, then it is used to classify the testing data, the most confident unlabeled records along with their predicted classes are added to the training set to re-train the classifier

(bootstrapping). The proposed technique had outperformed many other traditional techniques such as Neural Net, SVM, and Naïve Bayes in classifying customers in CRM to enhance its processes such as identifying valuable customers to retain or attract them.

In [5] a study on life insurance company CRM data was conducted to analyze the customer's data and to avoid customer churn\attrition. The authors had argued that a big data with a multi-class problem was tackled to classify the customer's willingness to continue or not. The dataset used in this study was extracted from operational database, unlike other insurance policies types, an agreement in life insurance is for an average of 18-20 years should be issued, so to build an efficient model the authors needed to mine data for a significant (big) period of time. In [5] paper the authors are interested in using demographic data like gender, age, profession, etc. Term of policy, sum assured, premium, Agent, etc as the policy details was also used. The authors had begun this study by visualize the attributes to study dependency and correlation to select the relative attributes or the attributes to be combined. In addition to visualization Correlation-based Feature Selection (CFS), and Information Gain techniques were used. The ROC graph technique was used to evaluate the different classifiers ac-curacy. ROC indifferent to change in class distribution which is common in do-mains such as churn while the distribution of data is went towards one class label (most frequent).

The classifiers used to predict the classes are: J48 decision tree and ANN with a standard Multilayer Perceptron using BP. Many issues apart from the evaluation of the classifiers were addressed, such as the huge number of attributes in the dataset, an efficient feature selection technique, maybe probabilistic one, could be used to solve this issue. In addition human interference still required in different phases of the study.

In the research of [6] the partitioning around Medoids clustering algorithm K-Mediods, was implemented on Telecom CRM post sales dataset which is stored in Teradata environment, to the purpose of segmenting customers' behavior while selling new products. The K-Medoids uses the most centric object in a cluster to represent the cluster instead of the mean (K-Means) that may not be-long to the cluster. This makes K-Medoids more robust than K-means which out-perform with high number of Ks. Customer's preferences such as: age, contract type, quantity sold, used media, customer area importance, department, and billing history were used to define the segment. The results of the study had shown that K-Medoids clustering algorithm is very efficient in large datasets such as CRM.

In a recent paper [7] the authors tried to empirically compare data mining methods: decision tree and logistic regression to build customer churn model. The authors have found that decision tree outperformed the logistic regression. The analysis was built using two different customers' data sets (15,519 and 19,919 customers) from UK operator mobile telecommunication data. The data set has seventeen variables or dimensions including demographic data, services used, services usage, cost of services, and marketing data. Three different decision tree algorithms (CART, C5.0 and CHAID) were used with accuracy of about 70% to predict if the customer will continue or not.

While [8] had been using customer telecommunication big data to build a framework for targeting not only important customer, but potential churn customers too. First the author used Recency Frequency Monetary (RFM) analysis technique to generate customers segments. Based on the common characteristics of each customer segments targeted marketing campaigns are de-signed. The dataset used is a combination of structured and unstructured data. The structure data includes the demographic data, number of minutes or messages, usage of internet…etc. while the unstructured data consists of customer feed-back, social media contents, location, downloaded applications, online purchasing data,… etc.

In [9] the author have been using twitter text (tweets) as a source of big data. The author succeeded to discover a number of local events and trending topics about Dubai during that period of time. The author have collected tweets for four months (136,000 tweets) to create a corpus. Then text mining with clustering techniques have been used by to conduct the experiment.

Within the next sections, we explore the insurance company dataset using clustering data mining techniques. We will start first by studying the characteristic of the dataset and its tendency toward clustering, we will implement two different clustering experiments using different conditions.

3. The Dataset

In this research we have been using THE INSURANCE COMPANY (TIC) 2000 dataset. This real-life dataset was published by Peter van der Putten, and owned by a Dutch data mining company Sentient Machine Research, Amsterdam.

This data constituted the CoIL Challenge 2000 data mining competition. TIC dataset was collected from real world Customer Relationship Management (CRM) data, and consists of 9,822 customer records, 5,822 record for training and the remaining for testing. Each record has 86 attributes, the first 43 attributes are rep-resenting customers' demographics, and the remaining 43 are representing customers' behavior or products ownership. All of the features have nominal values, with the last one (the target attribute for COIL 2000) being binominal [10]. TIC 2000 dataset is available on: TIC 2000 homepage: http://www.wi.leidenuniv.nl/~putten/library/cc2000/,and Edinburgh University http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html.

4. Background

4.1. Determining the clustering tendency of data

Before starting the experiment and applying any clustering technique on this dataset, we have to study the tendency of the data to have clusters or similarity be-tween objects. In big datasets, such as the one we have with high dimensions, the curse of dimensionality could have a critical effect on the similarity measure. To reveal the tendency of the data to be clustered we had plotted the histogram of the pairwise distances of all objects in our dataset. If the resulted graph contains two peaks, this means that the dataset contains clusters. One of the peak to rep-resent the distance between objects in clusters, and the second peak to represent the average distance between objects as shown in Fig. 1. (a) [11]. The histogram of our data set as shown in Fig. 1. (b), is showing that the data has the tendency to be clustered.

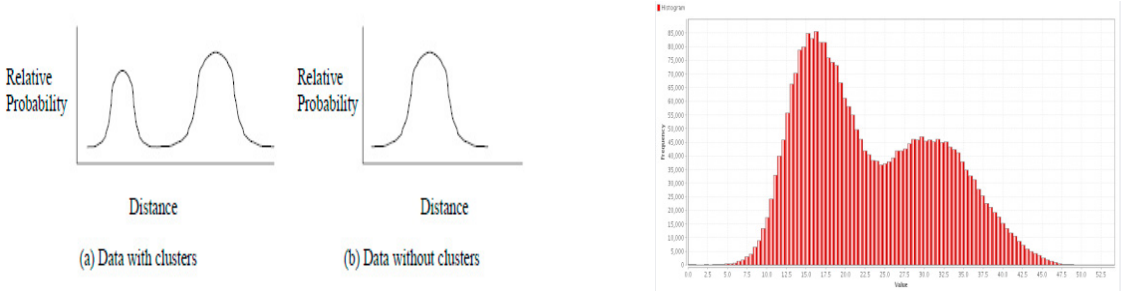


Fig. 1. (a) Plot of inter-point distances for data with and without clusters adapted from [11]; (b) Analysing Clustering Tendency-Distances Histogram for the COIL 2000 dataset.

4.2. K-Means Algorithm

K-Means is a partitioning clustering algorithm, where each cluster is connected with a centroid or central point (mean of points). During training each object assigned to a cluster with the closest centroid, usually Euclidean distance is used. Number of clusters K should be defined at the beginning and initial centroids are defined randomly. K-means Algorithm is shown in Fig. 2 [12].

ALGORITHM 13.1. K-means Algorithm

K-MEANS $(\mathbf{D}, k, \epsilon)$:

$t = 0$
Randomly initialize k centroids: $\boldsymbol{\mu}_1^t, \boldsymbol{\mu}_2^t, \ldots, \boldsymbol{\mu}_k^t \in \mathbb{R}^d$
repeat
  $t \leftarrow t + 1$
  $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$
  // Cluster Assignment Step
  foreach $\mathbf{x}_j \in \mathbf{D}$ **do**
    $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \boldsymbol{\mu}_i^t \right\|^2 \right\}$ // Assign $\mathbf{x}_j$ to closest centroid
    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$
  // Centroid Update Step
  foreach $i = 1$ *to* k **do**
    $\boldsymbol{\mu}_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$
until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

Fig. 2 K-means Clustering Algorithm, adapted from [12].

To define the best number of Ks we had been following the Elbow technique as discussed in the next section.

4.3. Elbow method to Choose Number of Clusters (K)

Elbow criterion is a way to define the best K in K-means clustering technique. As in Elbow technique, to define the best K we had been repeating the experiment 10 times with different number of clusters, and in each time we had been plotting the Average within Centroids Distance for each K, then we picked the K where the graph has an angle that follows by a drop and then no variance. At this angle is the best K or number of clusters for this dataset. The best K is 5 for the first clustering experiment, while it is 6 for the second one as shown in Fig. 3. (a) and (b) (The Data Science Lab, 2013).

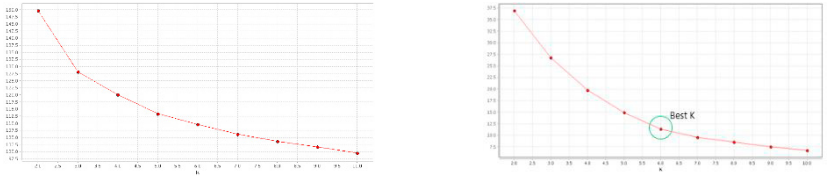


Fig. 3. (a) Elbow Method to determine the no of Ks- first Experiment; (b) no of Ks Second Experiment

4.4. Self-Organized Maps (SOM)

For a high dimensional dataset, visualization is challenging. To simplify the presentation and explore meaningful relationships we had been using the Self Organized Maps (SOM) or Kohonen Maps. SOM are neural networks that converts multidimensional data into two dimensional data representing the relationships between data objects. The location of the nodes on the map represents the similarity (order) to its neighbor in the feature space. So by reducing the high dimensions to a map, visualization becomes easy and attractive, at the same time grouping similar data together is a mean of clustering [13].

SOM is a special type of Neural Network that uses competitive learning to respond to the samples. For a sample vector, weights from the same size of the output network (no. of nodes) are randomly defined, then the Euclidean distance (commonly used) between the sample and weights should be calculated. The node with the minimum distance is the winner, this winner is considered to represent a cluster of similar objects or neighborhood. Next is the adaptation phase where weights of all neighborhoods nodes should be adjusted. “Learning rate should be decreasing

function of training epochs" [13].The adaptation will lead the weights to move toward the input attributes values, so it becomes more adapted to cluster similar records.

The variation in the color scheme from red to dark blue is showing the average Euclidean distance between adjacent nodes. The net shown in Fig. 4 called the SOM grid or map, this grid can be used to understand or read the dataset distribution. In this graphical representation red areas are representing dissimilarity (large Euclidean distances between objects). While the dark blue areas are representing similarity (small Euclidean distances between objects). Then the color degrade between orange to blue showing the reduction in distance between nodes. We also can use this color scheme to visualize each feature or attribute, so these colors will represent the value of each individual attribute (dark blue for low values, red for high values). The representation of individual attributes using this coloring schema will generate a grid called feature plane, as shown in Fig. s 26 to 35 (Appendix Section). (http://www.viscovery.net/self-organizing-maps).

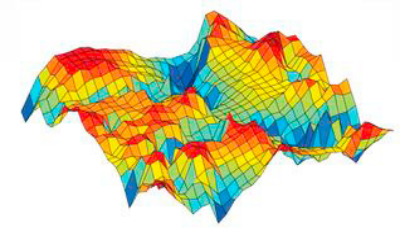


Fig. 4. Adapted from http://www.viscovery.net

In the following two sections we will executed two different experiments with two different techniques. We will show the execution details of each experiment, the evaluation, and we will analyze the results.

5. Experiment 1: Clustering by K-means

The first experiment to be executed is using K-means algorithm.

5.1. Execution

To run this experiment we had been using the most 21 informative attributes to the target attribute (caravan policy), then we had applied the Clustering (K-means) operator on the reduced dataset as shown in Fig. 5. As recognized from the table of centroids in Fig. 6, cluster 0 and cluster 2 are very near to each other. So they could share many characteristics as what we will see in the next sections.

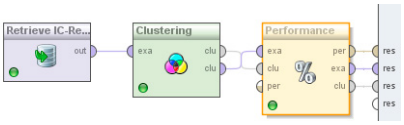


Fig. 5. Rapidminer process

Attribute Description	Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Customer Subtype	1-MOSTYPE	8.479	34.165	7.077	35.704	23.403
Customer main type	5 MOSHOOF	2.444	7.933	2.020	8.413	5.192
High level education	16 MOPLHO	1.788	0.539	2.930	0.976	1.449
High status	19 MBERHO	2.003	0.911	3.553	1.607	1.451
Social class A	25 MSKA	1.635	0.639	3.162	1.425	1.226
Rented house	30 MHHUUF	6.330	7.153	1.172	2.249	6.955
Home owners	31 MHKOOP	2.682	1.857	7.838	6.756	2.055
1 car	32 MAUT1	6.313	5.693	6.778	6.013	5.284
No car	34 MAUT0	1.944	2.547	0.988	1.624	3.277
National Health Service	35 MZFOND	6.135	7.483	4.634	6.320	6.934
Income < 30.000	37 MINKM30	2.692	3.942	1.108	1.946	3.972
Income 45-75.000	39 MINK457	2.681	1.586	4.111	2.947	1.974
Average income	42 MINKGEI	3.694	2.921	4.934	3.955	3.078
Purchasing power class	43 MKOOPK	5.540	3.067	6.773	3.766	2.112
Contribution private third party insurance	44 PWAPAR	0.682	0.711	0.854	0.710	0.951
Contribution car policies	47 PPERSAI	2.919	3.083	3.120	2.956	2.651
Contribution fire policies	59 PBRAND	1.560	1.546	2.205	2.116	1.288
Contribution boat policies	61 PPLEZIEI	0.019	0.012	0.027	0.020	0.015
Number of private third party insurance	65 AWAPAR	0.355	0.382	0.438	0.371	0.495
Number of car policies	68 APERSAI	0.557	0.577	0.593	0.565	0.492
Number of boat policies	82 APLEZIEI	0.004	0.004	0.010	0.006	0.005

Fig. 6. Centroid Table

5.2. Evaluation

The results of the first experiment, K-means on a reduced dataset, had shown the following results. The value of Davies Bouldin is very small, means the intra-distance (between points in the same cluster) is very small and the inter-distance (between clusters) is very big, this shows a good clustering. In our experiment the value of Davies Bouldin is 1.632 the smallest for K=5.

5.3. Analysis

To read the clustering we had visualized the clusters, attributes, and the target class as a scatter plot for the most informative 10 attributes. All these scatter plots are shown in the appendix section. We had analyzed the resulted plot and had recognized the following interesting results:

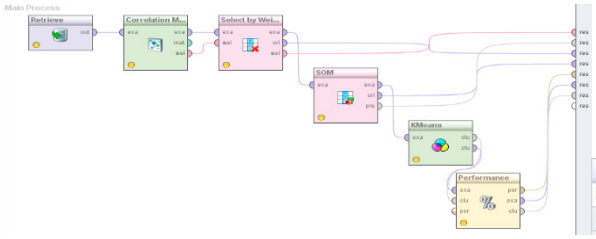
- Cluster A (C0): Singles, families with Adults, Seniors, Retired, and Religious Farmers in this cluster are more likely to subscribe in a caravan Policy. Because a caravan is made of wood, customers in this cluster usually have 3-4 contribution in fire policy.
- Cluster B (C1): Business Men and Retired, and Cruising Seniors customers tend to do not own a caravan policy. This could be explained by the nature of these families of customers, for example cruising seniors would prefer a sea trip instead of camp out in a caravan, at the same time their low purchasing power and average income could be a reason.
- Cluster C (C2): Customers in this segment are Living well singles, average families, or Business Men. People in this segment are more likely to own a Caravan Policy according to their Intermediate to High Purchasing Power and Average Income.
- Cluster D (C3): This cluster includes Seniors, Retired, and Religious Farmers. These people with 4-7 car policy contributions are more likely to subscribe in a caravan policy. This is logically because any caravan needs to be pulled by a car.
- Cluster E (C4): While Seniors, Retired, and Religious Farmers with the max contribution in boat policies have higher tendency to own a caravan policy from other customers. Also customers with 2 contributions of private third party insurance are more likely to own a caravan policy.

6. Experiment 2: Self Organized map (SOM) and K-means

In this experiment we will combine both K-means and Self Organized Map (SOM).

6.1. Execution

We had been using Rapidminer 5.3 to conduct the experiment as shown in Fig. 7. We had started the experiment by avoiding duplicate in behavior and to do this we had studied the correlation between attributes and ignored the associated (High Weights) ones. After that we had been using the SOM technique for not only the sake of reducing dimensions, but also for visualizing the clustering in a readable, easy, and fast way. Then we had re-clustered the features resulted from SOM using K-means algorithm.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
SOM_0	16.905	13.560	23.955	2.047	23.012	1.947
SOM_1	12.462	1.268	11.837	2.476	3.198	10.813

Fig. 7. (a) Rapidminer Process; (b) SOM Centroid Table

The SOM model generated by the experiment is shown in Fig. 8. This model is showing how SOM can visualize high dimensional dataset and the relation between samples in 2D map.

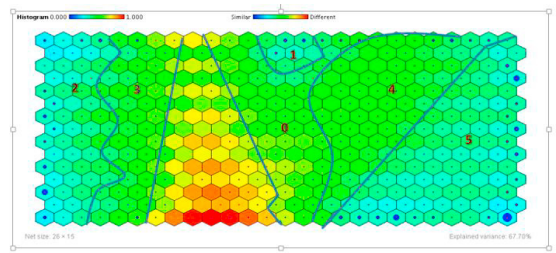


Fig. 8. SOM

The following figure is showing the scatter plot of the new 2D feature space resulted from SOM within the 6 resulted clusters.

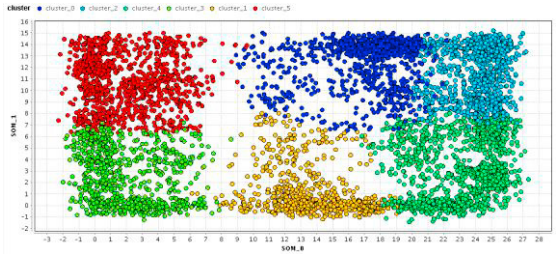


Fig. 9. SOM dimensions scatter plot.

6.2. Evaluation

We had been using Davies Bouldin to assess how well the results of a cluster analysis fit the data without reference to external information (IG). Low Davies Bouldin Index means that intra-distance is very small and the inter-distance is very high leading to an ideal clustering. For our experiment the value of Davies Bouldin measure is 0.699.

6.3. Analysis

To study the behavior of customers within each cluster we had been using the feature plane for each attribute as shown in the Appendix. The variation of color within the same feature is representing different values, dark blue for low values, to red for high values. By comparing the clustering map with the feature planes of the 10 most interesting attributes, we conclude our experiment with the following analysis of customers' demographics and behavior:

- Cluster A (C0): This is the cluster of Singles Religious Seniors. A customer belongs to this cluster has Higher average income than Cluster B, high purchasing power, high contribution for car policies , high contribution of private third party insurance.
- Cluster B (C1): Is the cluster of Living well Seniors. Customers in this cluster have a high average income and purchasing power, but lower than Cluster A, a high contribution for car policies, a high contribution of private third party insurance.
- Cluster C (C2): The customers in this segment are Middle Class Provincials Home Owners families. They have high income and purchasing power. They prefer higher contribution to fire policy and private third party insurance than Cluster D. They also prefer private health service more than national.
- Cluster D (C3): This the cluster of Affluent starting young families, with fair number of home owners, with intermediate average income and high purchasing power. They have low contribution of private third party insurance.
- Cluster E (C4): These customers spend less with higher income than Cluster F, and have high contribution private third party insurance.
- Cluster F (C5): It is the segment of Large Families, Employed child, Village families, Couples with teens, mixed small town dwellers, or Traditional families. In this segment customers spend more with lower income than Cluster E.

By using the SOM visualization in addition to the feature planes reading, any insurance company would be able to understand its customer's demographics and behaviors in a big dataset. Using this analysis in a cross-marketing campaign to offer the customers new policies could be done as follows:

- Customer, who prefers National Health Service, insurance packages with national hospitals or clinics.
- Customer, who prefers private health service than national, an insurance packages with private hospitals or clinics.
- Customer, who spends less with higher income than other clusters, a life insurance policy.
- Customer, who spends less with higher income than other clusters, and has no caravan policy, but has car and fire policies, a caravan policy.
- Customer, with complete family and high to intermediate income and have car insurance, accidents insurance policy.
- Customers, who are home owners, property insurance policies.

6.4. Results

After implementing the above proposed solutions on TIC dataset, the following results are obtained:

Table 1. Experimental Results

Parameters	K-Means	SOM With K-Means
No. Of Clusters	5	6
Execution Time	16 seconds	4 seconds
Davides Bouldin	1.632	0.699
Visualization	Difficult to visualize clusters	Clusters can be recognized easily

As recognized from Table 1, using SOM along with the K-Means algorithm to cluster the dataset had outperformed using K-Means alone in many dimensions. The two experiments were conducted on the same PC (Intel Core i7 2.10 GHz, 8.0 GB RAM).

7. Conclusion & Future work

To conclude there are always interesting information that could be explored in customers' data or CRM. This domain is not yet investigated well by data mining research, most of the data have never been analyzed and sometimes not even automated. At the same time and because of CRM dimensions human experts usually take longer time to analyze this data with less accuracy. Also there are many effective data mining techniques to study CRM datasets, we have been examining and testing some of them within this paper. Using those advanced techniques in parallel with the data mining consultant would improve businesses such as Insurance.

One of the key purposes of marketing is to detect customer's characteristics and analyze it by segmentation. Setting marketing strategies and campaigns would be more effective using the resulted demographic or behavioral segments instead of using the same marketing plan for all customers. To achieve this goal, we proposes 3 different solutions in this paper based on different data mining techniques and THE INSURANCE COMPANY (TIC) 2000 CRM dataset.

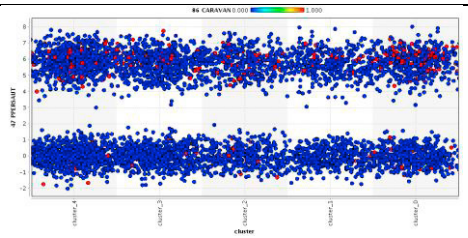
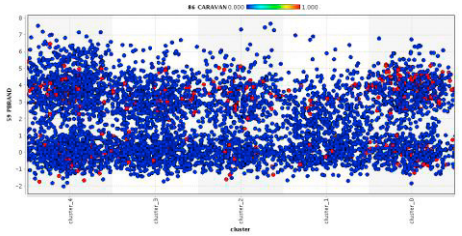
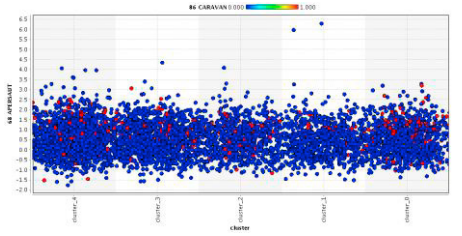
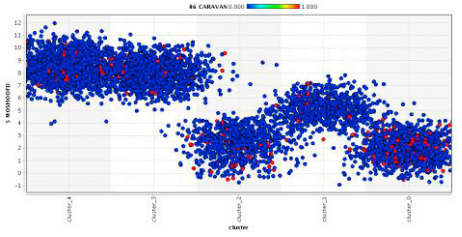
Starting from using IG with K-means to explore the characteristics of customers who are likely to buy a Caravan policy in our first solution. Studying these characteristics (both demographic and behavioral) could be useful in understanding the market and designing cross-marketing Campaigns.

While in the second experiment we had been focusing on visualization in CRM using SOM method, at the same time the experiment had shown promising results for clustering using K-means. Because of the dataset tendency toward clustering, it is interesting to have a general understanding of customers, their characteristics and their needs. Reading these information in such a high dimensional dataset is really challenging, SOM had shown interesting results, in the second proposed solution. At the same time the experiment had shown a good clustering quality by recording a very small Davies Bouldin value compared with the first clustering experiment. This understanding of the customer's behavior and demographics is important for customer-centric businesses.

Trying different data mining techniques and methods to study the demographic features and behavior of a customer in any CRM data set is really interesting. After using these various methods it is very clear to us it is how we look into the data not only what technique we use. The intuition and vision of the analyst are the spirit of data mining, the ability to find insights or interesting patterns in any data set regardless of its dimensions, complexity, or even how accurate is the algorithm used. As a future plan, extra efforts should be spent toward applying our proposed solutions on a novel data, especially in the Middle East, and in different domains, also studying the behavior features of a customer for a period of time could be effective. Also trying to predict new customer behavior or interest as who is interested in owning X insurance policy could be done. Finally integrating more than datasets related to customers' behavior such as their car accidents or health records could reveal different interesting knowledge in new dimensions.

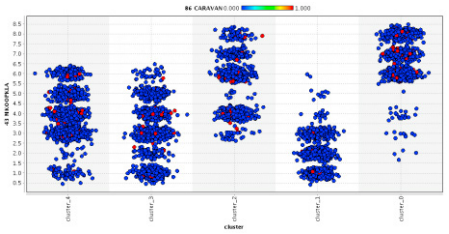
Appendix A.

A.1. Clustering Analysis

Attribute / Values	Cluster, attribute, and target class scatter plot	
47 PPERSAUT Contribution car policies 0 - 8		Customers with 0-2 car policy contributions has less chance to own a Caravan. Customers with **4-7 car** policy contributions, especially in **C3,** more likely to own a Caravan Policy.
	Fig. 10. Contribution Car Policies	
59 PBRAND Contribution fire policies 0 - 8		Customers with **3-4** contribution in **fire** policies, in **C0, C3**, are more likely to own a Caravan Policy. Few has more than 6
	Fig. 11. Contribution Fire Policies	
68 APERSAUT Number of car policies 0 – 7		Most of the customers have <= 4 car policies C0,C2,C3,C4 More likely to own a Caravan Policy
	Fig. 12. Number of Care policies	
5 MOSHOOFD Customer main type L2 1 Successful hedonists 2 Driven Growers 3 Average Family 4 Career Loners 5 Living well 6 Cruising Seniors 7 Retired and Religious 8 Family with grown ups 9 Conservative families 10 farmers		Customers in C0, C2 are from type < = 5 Customers in C3, C4 are from type > = 6 C1 has customers from 4-7 type, Career Loners, Living well, cruising seniors, and retired and religious families tend to do not own a Caravan policy. (Low Purchasing Power & Average Income) C0,C2,C3,C4 are more likely to own a Caravan Policy(Intermediate to High Purchasing Power & Average Income)
	Fig. 13. Customer main type L2	

43 MKOOPKLA
Purchasing power class

1 - 8



C0 > 5

C2 2-8 has more ones, purchasing power is >=6

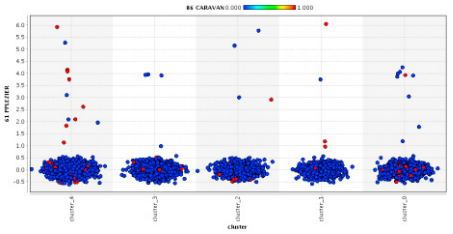
C3,C4 <=6

C1 <=4 Customers tend to do not own a Caravan policy.

Fig. 14. Purchasing power class

61 PPLEZIER
Contribution boat policies

0 - 6



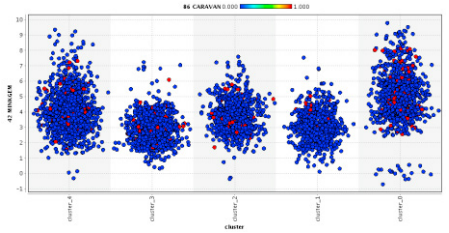
Max contribution in C4 with higher potential to own a Caravan Policy(compare with main type L2)

Most of Customers has 0 boat policy contribution

Fig. 15. Contribution boat policies

42 MINKGEM
Average income

0 - 9



C0,C4 have max income and more likely to own a Caravan Policy

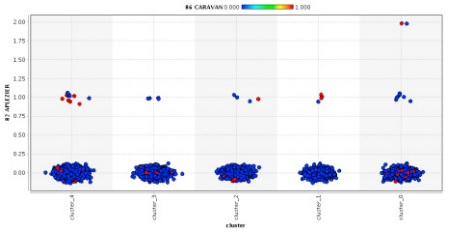
C0,C4 has >=7

C1,C2,C3 <=6

Fig. 16. Average income

82 APLEZIER
Number of boat policies

0 - 2



C0, C1, C2, C4 more likely to own a Caravan Policy when they own 1, or 2 boat policies.

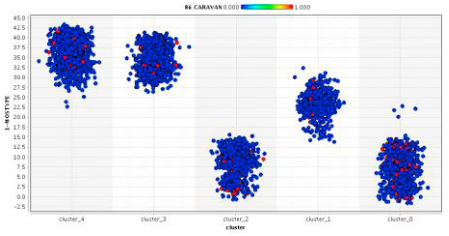
Max no in C4 with higher potential to own a Caravan Policy

All Clusters could have Caravan Policy contribution when no of boat policies=0, min in C1

Fig. 17. Number of Boat policies

1 MOSTYPE
Customer Subtype see L0

1 - 41



C0,C2 < 13, tend to have more Caravan policy owners(High Purchasing Power & Average Income)

C1 14 – 30 tend to have less Caravan policy owners.

C3, C4 25-40 more likely to own a Caravan policy (Intermediate to High Purchasing Power & Average Income)

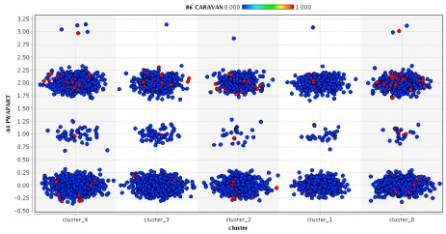
Fig. 18. Customer Subtype L0

44 PWAPART
Contribution private third party insurance see L4

0 – 9

Available

0-3



Most of the records 0, then 2 contributions.

The customer with 2 contributions more likely to own a caravan policy.

At the same time customer with 0 contribution showed a tendency toward owning a Caravan Policy.

C4 has the most tendency.

Fig. 19. Contribution private third party insurance L4

A.2. Feature Planes (only first 10 informative)

Contribution car policies 0-8

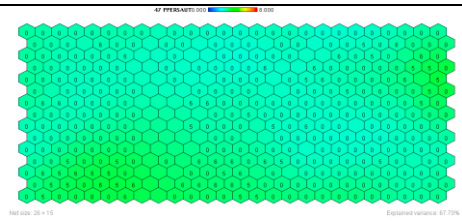


Fig. 20. Feature1 Plane

Contribution fire policies 0-8

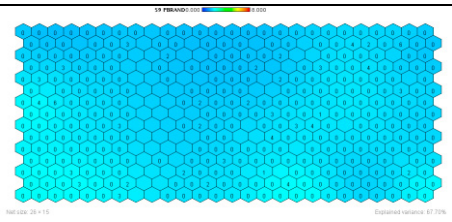


Fig. 21. Feature2 Plane

Number of car policies 0-7

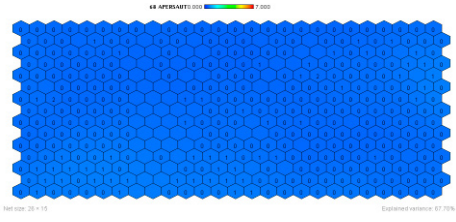


Fig. 22. Feature3 Plane

Customer main type see L2

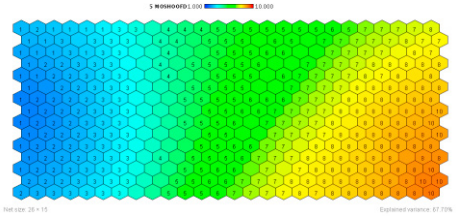


Fig. 23. Feature4 Plane

Purchasing power class

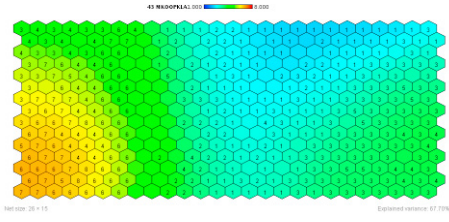


Fig. 24. Feature5 Plane

Contribution boat policies

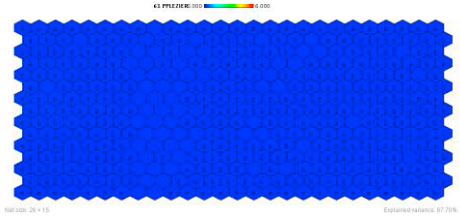


Fig. 25. Feature6 Plane

Average income

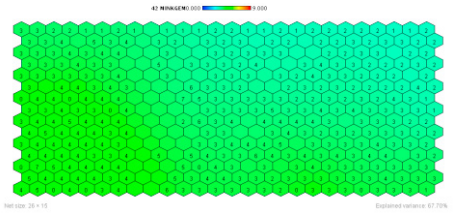


Fig. 26. Feature7 Plane

Number of boat policies

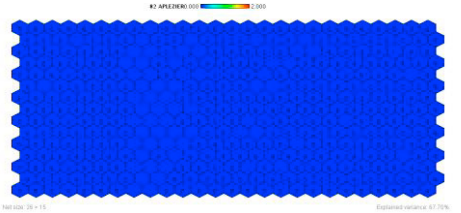


Fig. 27. Feature8 Plane

Customer Subtype see L0

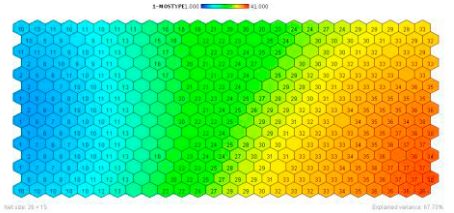


Fig. 28. Feature9 Plane

Private third party insurance see L4

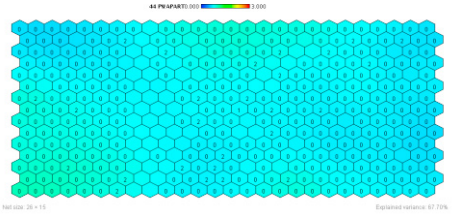


Fig. 29. Feature10 Plane

References

[1] Namvar, M., Gholamian, M. R. and KhakAbi, S. (2010). A two phase clustering method for intelligent customer segmentation. In Intelligent Systems, Modelling and Simulation (ISMS), 2010 International Conference on (pp. 215-219). IEEE.

[2] Khajvand, M. and Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327-1332.

[3] Zadeh, R. B. K., Faraahi, A. and Mastali, A. (2011). Profiling bank customers behavior using cluster analysis for profitability. In International Conference on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.

[4] Ahuja, V. and Medury, Y. (2011) "Corporate blogs as tools for consumer segmentation-using cluster analysis for consumer profiling" *Journal of Targeting, Measurement and Analysis for Marketing* **19** (**3**): 173-182.

[5] Goonetilleke, T. O. and Caldera, H. A. (2013) "Mining Life Insurance Data for Customer Attrition Analysis" *Journal of Industrial and Intelligent Information* **1** (**1**).

[6] Mbarki, J. and Jaara, E. M. (2014). Deployment of Partitioning Around Medoids Clustering Algorithm on a Set of Objects Derived from Analytical CRM Data.

[7] Hassouna, M., Tarhini, A., Elyas, T. and AbouTrab, M.S., (2016). Customer Churn in Mobile Markets A Comparison of Techniques. arXiv preprint arXiv:1607.07792.

[8] Singh, I. and Singh, S. (2017) "Framework for targeting high value customers and potential churn customers in telecom using big data analytics" *International Journal of Education and Management Engineering* **7** (**1**): 36-45.

[9] Hamadeh, M.W. and Abdallah, S. (2017). Discover Trending Topics of Interest to Governments. In International Conference on Advanced Intelligent Systems and Informatics (pp. 366-373). Springer, Cham.

[10] P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company Case.

[11] Steinbach, M., Ertöz, L. and Kumar, V. (2004). The challenges of clustering high dimensional data. In New Directions in Statistical Physics (pp. 273-309). Springer Berlin Heidelberg.

[12] Zaki, Mohammed J., and Wagner Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge UP, 2014. Print

[13] Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.