Research Journal of Applied Sciences, Engineering and Technology 7(4): 786-790, 2014
DOI:10.19026/rjaset.7.317
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: April 06, 2013 Accepted: June 25, 2013 Published: January 27, 2014

Research Article

Deployment of Partitioning Around Medoids Clustering Algorithm on a Set of Objects Derived from Analytical CRM Data

J. Mbarki and E.M. Jaara
Laboratory of Computer Science Research (LARI), Faculty of Sciences,
University Mohamed Ier, Oujda, Morocco

Abstract: The aim of this study is to highlight the importance of the unsupervised learning in Data mining and CRM fields. Data mining commonly known by its acronym KDD: knowledge discovery in data base, it refers to all methods and algorithms used for data exploration or prediction in large data bases volumes, Data mining is very important in various fields such as science, business and other areas deal with a large data set. CRM: Customer Relationship Management is an integrated information system that is used to plan, schedule and control the pre-sales and post-sales activities in an organization, both CRM and data mining techniques helps organizations maximize the value of every customer interaction and drive superior corporate performance. Clustering is one of the favoured used methods in data mining: The objective of this study is to implement the clustering algorithm K-Medoids via a shell script applied on a set of Analytical CRM data stored in Teradata environment.

Keywords: Clustering, CRM, data set, database, teradata environment

INTRODUCTION

Data Ware House (DWH): "Data warehouse is a subject oriented, integrated, non-volatile and time variant collection of data in support of management's decisions "(Inmon, 1996), It's a large reservoir of detailed and summary data that describes the organization and its activities, repartitioned into a various business dimensions. The customer dimension remains the most challenging dimension within DWH.

CRM: Customer relationship Management is the process of managing all aspects of interaction a company has with its Customers, with one of its most important dimensions being Analytical CRM. Analytical CRM main's functions are to enable storage and analysis of data generated by such as operations, marketing, sales, customer service and even information Collected about customer during the contact with the latter.

The wealth of any business relies heavily on customers: because the main business objective of any company is to Increase profitability and customer loyalty by:

- Controlling risks
- Using the right channels at the right time

This is why it's important to better know the customer and their behaviours and this is why customer relationship management must be implemented as the primary company philosophy and the strategies it promotes have to be adopted within the company. The Global objective of this study is to implement a data mining process in telecom CRM post sales Data, in order to help study the impact of customer segmentation and the used channel (indirect via partners, direct via the company shops or via internet media) and hence to take the appropriate decisions accordingly, because the close monitoring is pretty helpful when embarking on any new product selling.

Data mining: It's the process to find useful patterns from a data in a large database (Fayyad *et al*., 1996), Data mining uses a variety of techniques derived from statistics, machine learning, artificial intelligence and database technology. The techniques are grouped into two broad categories (Descriptive and predictive ones). Classification and clustering are the main examples representing respectively the two sets.

Clustering: It's the unsupervised learning task which aims to arrange the instances described in a database into a set of homogeneous and mixed clusters. Clustering algorithms fall into two distinct types: Hierarchical and partitioning methods (Fraley and Raftery, 1998). Additional classification has been introduced later (Han and Kamber, 2001) with three new layers: Density-based methods, model-based clustering and grid based methods. The main purpose of

Corresponding Author: J. Mbarki, Laboratory of Computer Science Research (LARI), Faculty of Sciences, University Mohamed Ier, Oujda, Morocco

this study is to deploy the partitioning around Medoids algorithm on two sets of sales data: These 2 sets year: "Smart Home Services"; This product offers a wide variety of remote services at home, it allows for example Mobile users to have a continuous view on their home and to be notified in case intrusion and the second set is "Money services on Mobile product": this innovative product provide remote mobile money services such as: Retail payments, Money transfers, E-commerce, Savings, transactions, Bill payment, Salary disbursement (on-going project), parking payment, Microfinance… and more.

K-Medoids algorithm: K-Medoids algorithm is one of the well-known clustering algorithm, called also in its first version (Partitioning around Medoids- (Kaufman and Rousseeuw, 1987)). This method is similar to the K-means algorithm. It differs mainly in its representation of the different clusters: Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster.

The advantage of PAM compared to K-means, is that it is more robust but its disadvantage is that it's less performant in case the number of partition k and total count n are high, in this case Complexity is due to the Global Cost calculation. Both methods require the user to specify K, the number of clusters. Below is the technical principle of the algorithm:

Begin

(1) Randomly select k as representatives of X_1,….. X_k classes represent sales realised during the first quarter of 2013 of respectively two new products, launched end of last

(2) Calculate the Global Cost of swap of each couple (X_i, O_h) where X_i a class representative and O_h another point
Let $CC_{ih} = \sum jd(i, h) - d(i, j)$ the global gained distance obtained by the swap of h by i

(3) Select the couple (X_i, O_h) where CC_{ih} is minimal

(4) If CC_{ih} is negative for a chosen couple then swap X_i and O_h roles and Go to (2)
Return classes related to X_1,…..X_k representatives, End

EXPERIMENTAL METHODOLOGY

Within the framework of the joint coordination and close ties among university Med1 and industry, the input of our process was sales data of a mobile telecom company.

The expected size of our input is relatively low, hence the choice of PAM clustering algorithm in place of the two other versions CLARANS (Clustering Large Applications based upon Randomized Search) or CLARA (Clustering Large Applications).

We implement the K-Medoids algorithm in Linux environment with access to Teradata, the complete components (Fig. 1) are:

- **Remote system:** A company development server; it's a Linux development server with alias name

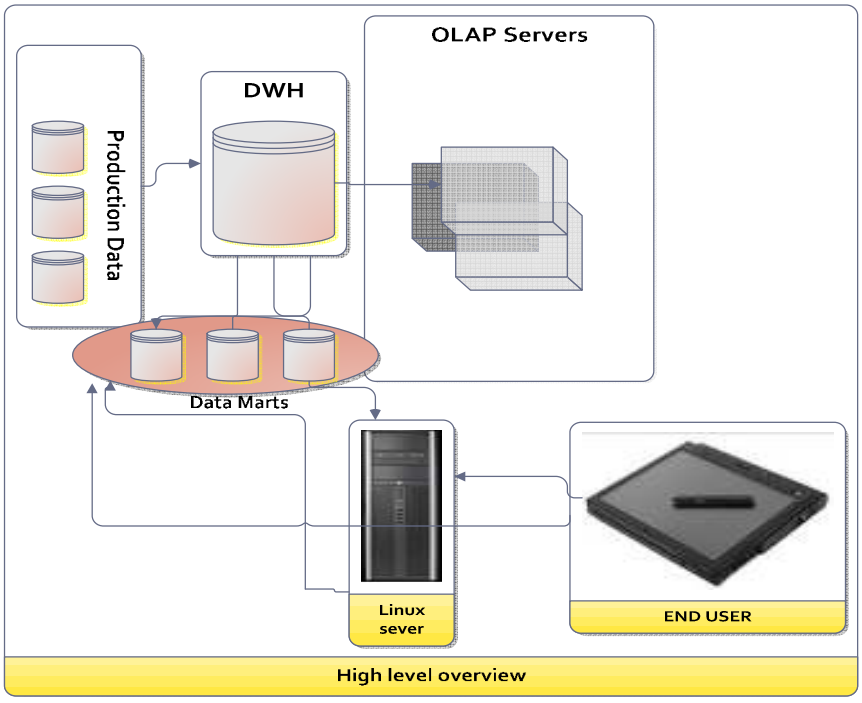


Fig. 1: High overview of our process components

equal to "aa799dev" and where the process can be created

- "Ultra-edit-32 professional Text/Hex Editor 9.20b" to create, to edit and to save the shell script, for our Experiment the script is called K_mod.sh it's building according to the K medoïds principle
- Access to Linux server where a home directory called univer_med1@aa799dev has been also created for us
- **Teradata SQL assistant:** The discovery tool to view and retrieve data from Teradata
- **Putty:** As user console to connect to the company Linux remote systems

For our experiments the dissimilarity measure is taken to be squared Euclidean distance D (xi, xi'). PAM like K-means generates different clusters indifferent runs (Murat *et al*., 2011).

Steps for creating and execution the PAM clustering algorithm:

Step 1: **Computational step:** Create the shell script in the LINUX server *pam.sh* according to clustering partitioning around Medoïd algorithm principle.

Step 2: Create-2 sets of data representing respectively sales of Smart Home product realised during the first quarter of 2013 and a second table containing Sales of Money services on Mobile product for Q1.

Step 3: Execute Fast load tool in order to import data for treatment.

Step 4: Execute the sh script and load data into tables.

Step 5: Go to Teradata playground and Select the 2 tables to view their contents.

EXPERIMENTAL RESULTS

Data Preparation is one of the most important steps to deploy Data mining processes. To implement our process we Create 2 tables representing respectively sales data of Money services product on mobile and those of smart home product; sales realised during the first quarter of this year (i.e., Execution_date< 01/04/2013), for the two sets inputs for our experiment. Each table is containing the following Field:

Cutomer_id: Identify uniquely the Mobile customer

Age: Calculated based on the customer date of birth (info retrieved from customer data base)

Product_id: Unique id of the product

Execution_date: Is the transaction date

Quantity_sold: Integer representing the acquired quantity

Used_media: Channel that executes transaction: three values are possible: "I" for internet, "Dir" for direct and "Ind"for indirect channel

Potentiel_degree_segment: This is a micro predefined segment assigned to the customer and dynamically maintained within matrix, taking into account a certain number of profile components like: contract type, customer area importance, department, billing history and age, three values have been defined (1, 2, 3) where 1 denotes potentially very important customer and 3 normal customer.

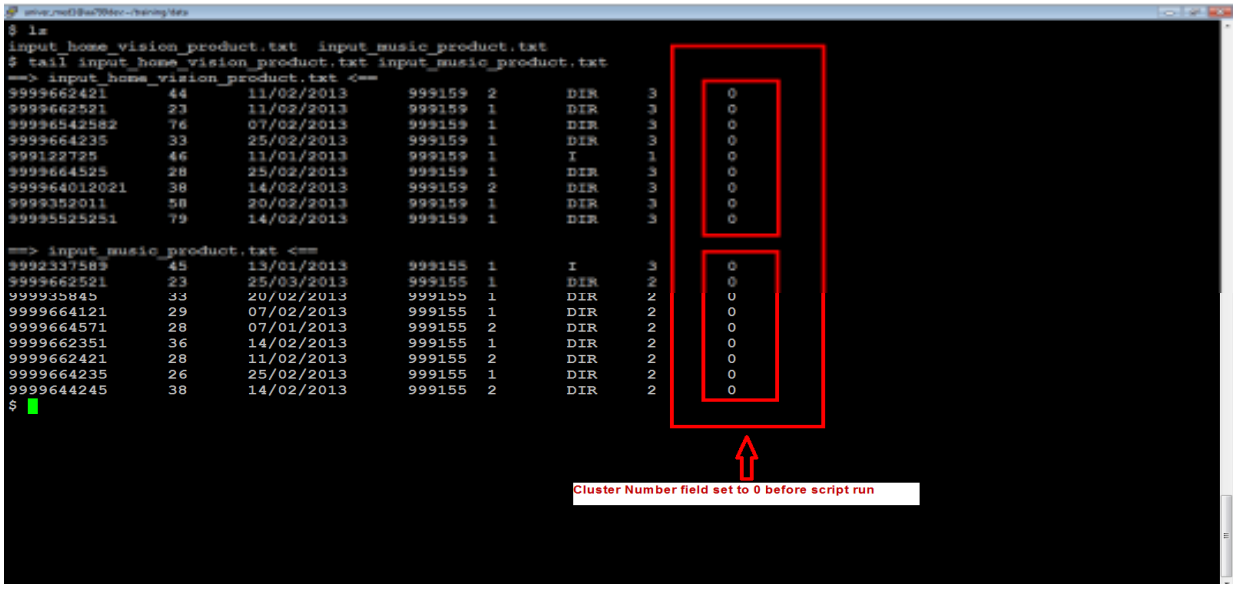
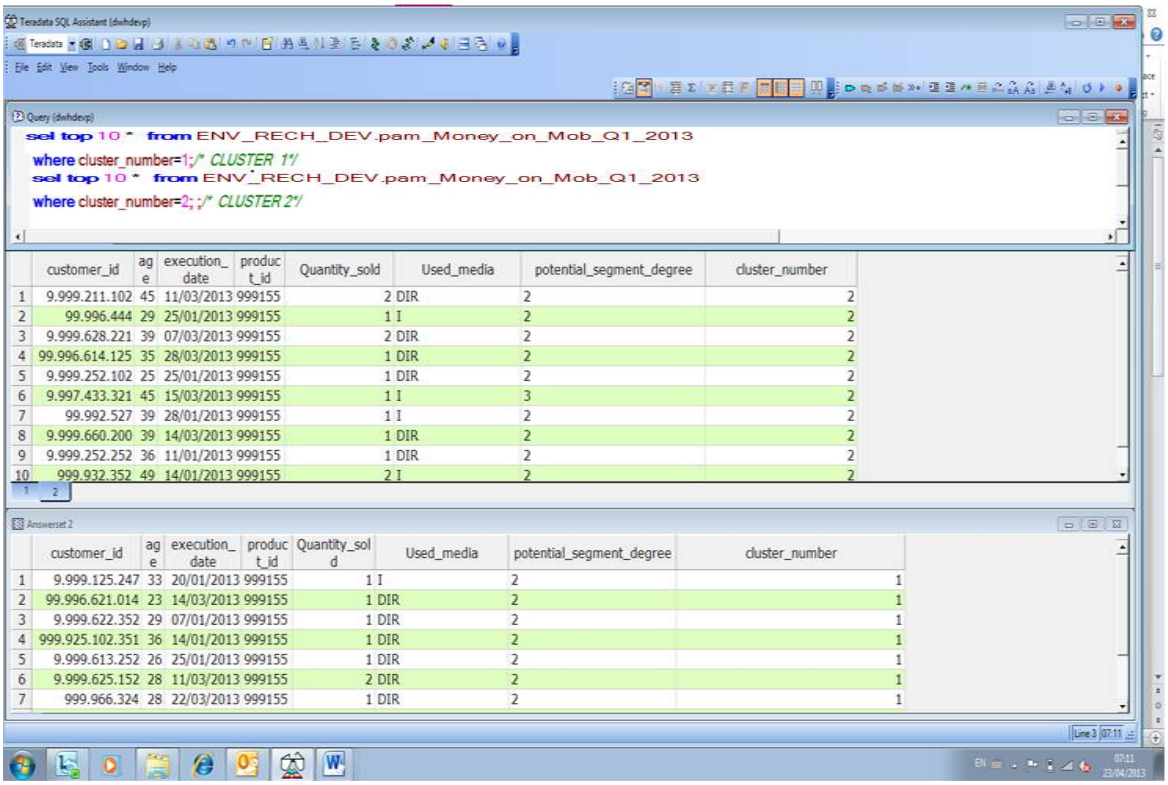


Fig. 2: Data input: view of the two sets (view of the last 10 records) where cluster_number = 0 (default value)



	customer_id	age	execution_date	product_id	Quantity_sold	Used_media	potential_segment_degree	cluster_number
1	9.999.211.102	45	11/03/2013	999155	2	DIR	2	2
2	99.996.444	29	25/01/2013	999155	1	I	2	2
3	9.999.628.221	39	07/03/2013	999155	2	DIR	2	2
4	99.996.614.125	35	28/03/2013	999155	1	DIR	2	2
5	9.999.252.102	25	25/01/2013	999155	1	DIR	2	2
6	9.997.433.321	45	15/03/2013	999155	1	I	3	2
7	99.992.527	39	28/01/2013	999155	1	I	2	2
8	9.999.660.200	39	14/03/2013	999155	1	DIR	2	2
9	9.999.252.252	36	11/01/2013	999155	1	DIR	2	2
10	999.932.352	49	14/01/2013	999155	2	I	2	2

	customer_id	age	execution_date	product_id	Quantity_sold	Used_media	potential_segment_degree	cluster_number
1	9.999.125.247	33	20/01/2013	999155	1	I	2	1
2	99.996.621.014	23	14/03/2013	999155	1	DIR	2	1
3	9.999.622.352	29	07/01/2013	999155	1	DIR	2	1
4	999.925.102.351	36	14/01/2013	999155	1	DIR	2	1
5	9.999.613.252	26	25/01/2013	999155	1	DIR	2	1
6	9.999.625.152	28	11/03/2013	999155	2	DIR	2	1
7	999.966.324	28	22/03/2013	999155	1	DIR	2	1

Fig. 3: Clusters view of sales data on money services product, since k = 2, value of cluster_ number field in (1, 2) view of top 10 records

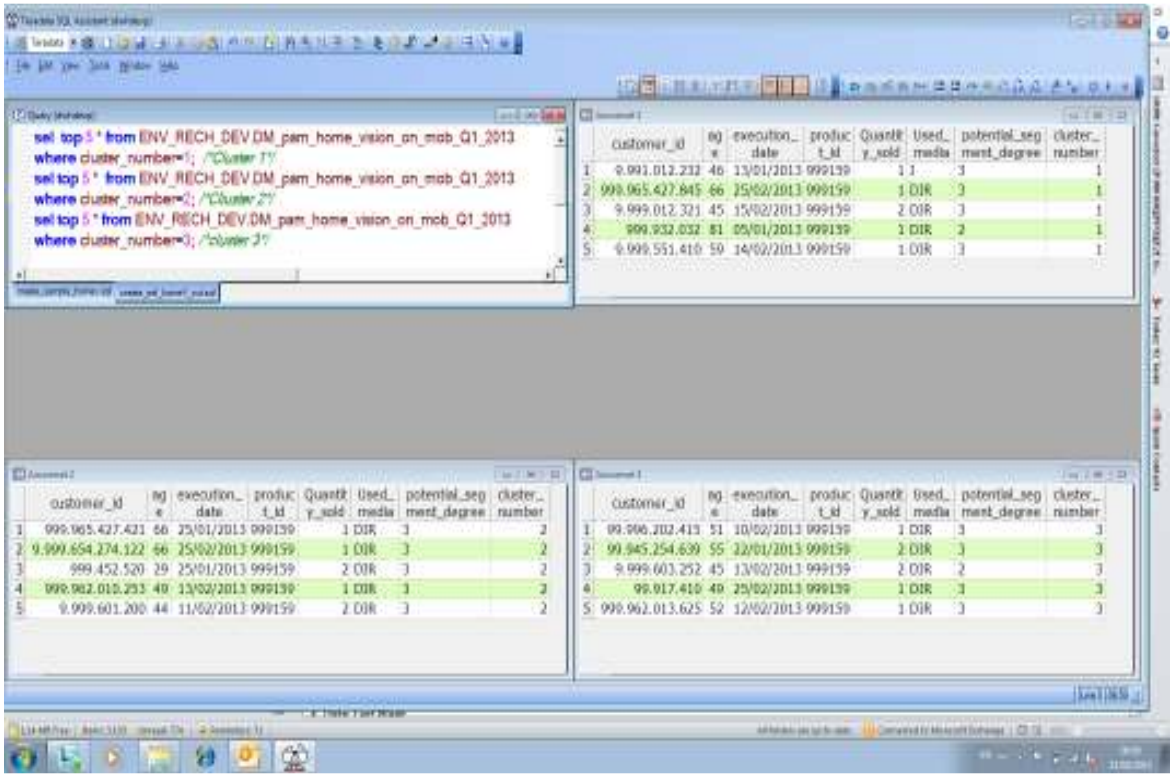


Fig. 4: Clusters view of sales data on smart home product, since k = 3, value of cluster_number field in (1, 2, 3): view of top 5 records

Example an SME customer from department (Ile de France with more than 100 active subscriber numbers is flagged to 1 whatever the age…

And finally an integer flag called cluster_number (this attribute is filled by K-Medoids algorithm K_mod.sh; this filed determine the cluster number to which Customer is belonging, the default value equal to 0 (value before script run) Fig. 2.

The value of k will be fixed in advance; k is determining the number of desired Clusters.

Two runs are foreseen for respectively k = 2 and k = 3.

For the first experiment data set was "Money Services", in this case the field prod_id = 999155 we assume that k = 2. Figure 3 shows sales data segmented into 3 clusters, the where clause in the select statement allows to view the clusters content.

The second data set used was sales on "Smart Home"; in this case the attribute prod_id = 999159; we assume that k = 3. Figure 4 shows the output clusters of sales data. Since the value of k = 3 the number of generated clusters is 3, the where clause in the select statement applied on the field cluster_number permits also to view the segments content.

CONCLUSION

It is widely recognised that clustering algorithms are very important algorithms of data mining to analyse real world situation. PAM algorithm is more efficient algorithm for mining large Databases and particularly data in CRM analytical world; it allows evaluating post-sales activities in an organization. So in this study, we focused the implementation of K Medoids in CRM data stored in Teradata environment and the experimental results demonstrate that it works well.

REFERENCES

Fayyad, U.M., G. Piatetsky Shapiro, P. Smyth and R. Uthurusamy, 1996. Advancesin Knowledge Discovery and Data Mining. AAAI Press/the MIT Press, pp: 573-592.

Fraley, C. and A.E. Raftery, 1998. How many clusters to which clustering method? Answers viamodel-based cluster analysis. Comput. J., 41: 578-588.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufman, Boston.

Inmon, W.H., 1996. Building the Data Warehouse. Wiley and Sons, NY.

Kaufman, L. and P.J. Rousseeuw, 1987.Clusteringby Means of Medoids. In: Dodge, Y. (Ed.), Statistical Data Analysis based on the L1Norm. Elsevier/North Holland, Amsterdam, pp: 405-416.

Murat, E., C. Nazif and S. Sadullah, 2011. A newalgorithm forinitialcluster centersin k-meansalgorithm. Pattern Recogn. Let., 32: 1701-1705.