# Applying Data Mining Techniques in CRM

Ahmed Bahgat El Seddawy
Arab Academy for Science and Technology
College of Management
Ahmedbahgat2@hotmail.com

Dr. Ramadan Moawad
Arab Academy for Science and Technology
College of Computer Science
Rammoawad@yahoo.com

Dr. Maha Attia Hana
Helwan University
Faculty of Computers & Information
Mahana_eg@yahoo.com

## ABSTRACT

*Customer relationship management "CRM" is very important factor in enhancing the organization competitiveness. In this paper, Data mining "DM" techniques are used to improve customer services in a radiology centers. Clustering customers is needed to find unsatisfied need, promote services packages and create new service packages. The proposed system radiology data mining system "RDMS" consists of three components; preprocessing, clustering and post processing. The data collected is for a period of four month for 6700 transaction. Three data sets are constructed from the original data set by dividing the whole data into 90%, 85% and 80% for training and 10%, 15% and 20% for testing respectively. Three K-means model are used with k=10, 15 and 18 cluster and each data set is used to calibrate and test the model for a total of nine ones. It is found that the best model is the one with 15 clusters. The clustering results are represented to a medical specialist who found that some results are reasonable and others go along with the center type and its policy.*

**Keyword:** CRM, DM, KD, DWH, K-Means, HW, SW, RDMS, GCRM.

## 1. INTRODUCTION

Nowadays, there is increase in competition among different commercial organization in Internet. The most valuable factor in commercial transaction is the customer; therefore interest increases in studying customer relationship management 'CRM'. This interest directs research in the area of IT towards knowledge discovery 'KD' specifically data mining 'DM'. DM is used to reveal existent relations among data that are uneasily discovered by human.

This paper aims to use DM techniques in the analysis of customer's data in radiology centers. It aims to discover the hidden relations about customers needs and to support the radiology's centers to better serve the customer. It also aims to support management with knowledge about customers' unseen needs, interest and preference.

The paper is divided into five sections; section two reviews related work and is divided into three parts, part one is about CRM, part two is about DM, and part three is about applying DM in CRM. Section three present the propose Radiology Data Mining System 'RDMS'. Section four illustrates the experiment. Section five illustrates the results. Finally, section six concludes the research.

## 2. RELATED WORK

### 2.1 CRM

CRM is the philosophy, policy and coordinating strategy connecting different players within an organization to coordinate their efforts in creating an overall valuable series of experiences, products and services for the customer [1]. It is a combination of policies, processes, and strategies implemented by an organization to identify its customer interactions and to provide means to track customer data. It involves the use of technology in attracting new and profitable customers, while forming tighter bonds with existing ones [2].

It is important to note that while most CRM consumers view it as software ‘solution’, there is a growing realization in the corporate world that CRM is really a customer-centric strategy for doing business supported by software [3].

CRM types are

1. Operational CRM provides support to front office business processes, including sales, marketing and service.

2. Analytical CRM analyzes customer data for a variety of purposes, such as design and execution of targeted marketing campaigns to optimize marketing effectiveness and design, and to execute specific customer campaigns, including customer acquisition, cross-selling, up-selling and retention.

3. Sales Intelligence CRM is very similar to Analytical CRM, but it is intended as a more direct sales tool. Features include the delivery of "alerts" to sales people based on analysis of such factors as customer drift, sales performance, good and bad, customer trends, customer margins, and campaign management.

4. Campaign Management Software is marketing-oriented CRM software that combines elements of Operational and Analytical CRM and allows campaigns to be run on an existing client base. Campaign Management is used to create personalized offers when it is prohibitively expensive to personally contact each client.

5. Collaborative CRM aims to get various departments within a business, such as sales, technical support and marketing, to share the useful information collected from customers' interactions.

6. Geographic CRM ‘GCRM’ is a customer relation management information system which collaborate geographic information system and traditional CRM.

There are two types of CRM applications; hosted and non-hosted ones. The hosted application is a web based one that enables CRM work for a company with distributed work locations and no access to a secure intranet connection. The non-hosted application is implemented in companies with strong IT infrastructure.

**2.2 DM**

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [3]. DM techniques are the result of a long process of research and product development [4]. The evolution of DM [5] is shown in table 1.

Table 1: The evolution by DM [5]

| Evolutionary Step | Business Question | Enabling Technologies | Product Providers | Retrospective, static data delivery |
|---|---|---|---|---|
| **Data Collection (1960s)** | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retrospective, static data delivery |
| **Data Access (1980s)** | "What were unit sales in New England last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retrospective, dynamic data delivery at record level |
| **Data Warehousing & Decision Support (1990s)** | "What were unit sales in New England last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retrospective, dynamic data delivery at multiple levels |
| **Data Mining (Emerging Today)** | "What’s likely to happen to Boston unit sales next month? Why?" | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Prospective, proactive information delivery |

There are several processes for applying DM:

1. Definition of the business objective and expected operational environment.

2. Data selection is required to identify meaningful sample of data.

3. Data transformation that involves data representation in an appropriate format for mining algorithm.

4. Selection and implementation of data mining algorithm depends on the mining objective.

5. Analysis of the discovered outcomes is needed to formulate business outcomes.

6. Representing valuable business outcomes.

Data mining consists of five major elements; to extract, to transform, and to load transaction data onto the data warehouse system, to store and manage the data in a multidimensional database system, to provide data access to business analysts and information technology professionals, Analyze the data by application software, and finally to present the data in a useful format, such as a graph or table.

DM techniques usually fall into two categories, predictive or descriptive. Predictive DM uses historical data to infer something about future events. Predictive mining tasks use data to build a model to make predictions on unseen future events. Descriptive DM aims to find patterns in the data that provide some information about internal hidden relationships. Descriptive mining tasks characterize the general properties of the data and represent it in a meaningful way. Figure1 shows the classification of DM techniques.


Figure 1: DM Techniques [5]

Association Rule is used to discover relationships between attribute sets for a given input pattern. [6] Define sequence discovery as "a sequential technique is a given set of sequences find the complete set of frequent subsequences".

Clustering is "the process of organizing objects into groups whose members are alike in some way" [7]. A cluster is therefore a collection of objects which are "similar" among them and are "dissimilar" to the objects belonging to another cluster. So, it deals with finding the internal structure in a collection of data, figure 2

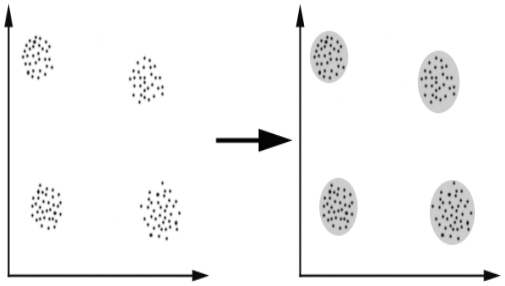

Figure 2: Simple graphical for clustering data [7]

[8] Define that "Clustering involves identifying a finite set of categories or segments 'clusters' to describe the data according to a certain metric". [9] Define that "Clustering enables to find specific discriminative factors or attributes for the studied data. Each member of a cluster should be very similar to other members in its cluster and very dissimilar to other clusters. When a new data is introduced, it is classified into the most similar cluster".

Several researchers classified clustering algorithms differently. Some classifies clusters as mutually exclusive, hierarchical or overlapping. Others classifies cluster into hierarchal and partitional. The most common classification for CRM application is shown in figure 3. Techniques for creating clusters include partitioning methods as in k-means algorithm, and hierarchical methods as in decision trees, and density-based methods.

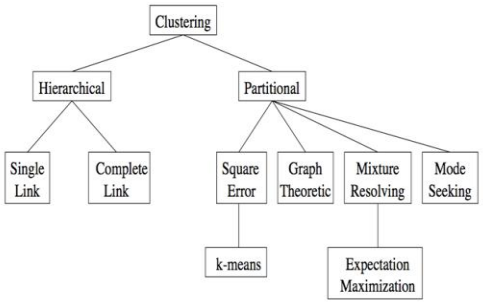


Figure 3: Clustering methods classifications for Moses Charikar [10]

### 2.3 CRM and DM Application

CRM is essential to compete effectively in today's marketplace. The more effectively customer information is used to meet their needs, the more the organization profit is. Operational CRM needs Analytical CRM with predictive DM models in order to build an effective model. The steps to build such as a model are:

1. Define business problem
2. Build or use marketing database
3. Prepare data for modeling
4. Build the model
5. Evaluate the model
6. Interpret the results

There are some common DM applications useful in the field of CRM. In retail sector, the use of store-branded credit cards and point-of-sale systems, enable to keep detailed records of every shopping transaction [13]. This enables better understanding for various customer segments. In bank applications, customer's transactions data is used to infer the customers' patterns and promote accordingly the bank services [14]. Telecommunication companies around the world face escalating competition which is forcing them to aggressively market special pricing programs aimed at retaining existing customers and attracting new ones [15].

## 3. THE PROPOSED RDMS

Radiology Data Mining System "RDMS" aims to build a data mining system for radiology centers in the medical sector. RDMS consists of three components; data preprocessing, data clustering and finally the post processing component.

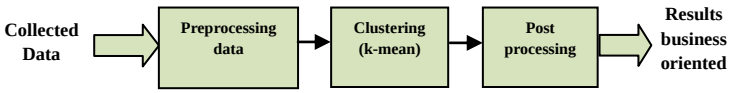


Figure 4: Show the RDMS process system

### 3.1 Preprocessing Data

Data preprocessing undergoes converting data from textual values to numeric, selecting main attributes for the system, converting data from numeric matrix to binary matrix and the last step is filtering data.

### 3.2 Clustering (K-Mean Algorithm)

[11] Defines K-means "as one of the simplest unsupervised learning algorithms". K-means steps are [12];

1. Assume number of cluster "K".
2. Pick cluster center at random
3. Calculate the distance between data sample and clusters.
4. Assign data sample to closest cluster center.
5. Compare calculate with previous cluster if result no repeat step 3 and 4, else end.

RDMS uses K-Means algorithm to segment Patients into groups with similar features. Figure 7 shows the flowchart of K-Means algorithm.

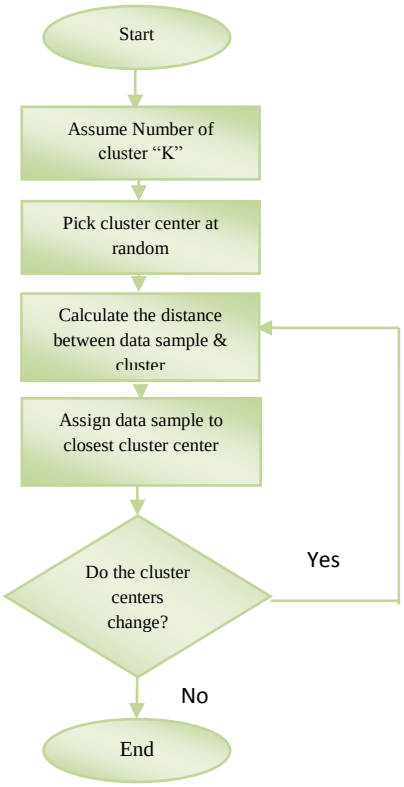


Figure 7: Show K-means work flow in process

### 3.3 Post Processing

The aim of this step is to visualize results in an easy way for the medical specialist to read and interpret. A quantified measure is used to representing output in a quantified form.

### 3.4 Computing Resources

Hardware for applying the RDMC system is a personal computer the configurations are Processor 3.2, Hard Disk 160 gaga, Ram 2 G and Monitor 17 Inch.

Operating system is windows XP services pack 3. Several software tools have been used. The first is Microsoft Excel sheets 2007 and has been used for analysis and filtering data. MatLab version 6.5 has been used in data preprocessing and data classification. The last software is the WEKA which is a collection of Java tools for DM written by staff at the University of Waiketo, New Zealand.

## 4. EXPERIMENTS

### 4.1 Data Collection

Data is from radiology center that is located in Egypt and has several branches. The center serves more than 10000 customers per year, contracting with more than 500 organizations and provides more than 450 scan type. All patients' data is stored electronically using SQL database. The center is willing to provide this research with recent patients' data from 1/1/2009 to 1/4/2009.

The database stores values for four fields; patient name, the employed organization, scans date and scan type as shown in Table 2. Table 3 indicates the format for each field. The number of data records is 6700 transaction for about 487 patients from 40 different organizations and those patients requested 30 scan types. The received data is in the form of excel sheet.

Table 2: Data

| Date | Patient Name | Scan type | Organization Name |
|---|---|---|---|
| 03/03/2009M | رامى | رسم قلب عادى | يونى كير TE-Data |
| 02/01/2009M | رامى | عادية على الصدر | يونى كير TE-Data |
| 01/01/2009M | فتحى نجيب | مسح ذرى على الكلى ( مرحلة واحدة ) | ت. صحى بنى سويف |
| 01/01/2009M | صادق امين | رنين على الفقرات القطنية | مؤسسة الأخبار |
| 11/04/2009M | كريم | دوبلر على القلب | نقابة الاطباء ( القاهرة ) |
| 01/01/2009M | فريد | عادية على المسالك | شركة ايجيكير |
| 01/01/2009M | فريد | تليفزيونية على البطن والحوض | شركة ايجيكير |
| 21/01/2009M | ناديه محمد | فحص ماموجرافى على الثدى الايسر | ت ص شمال الصعيد |
| 01/01/2009M | سميرة | رنين على الفقرات القطنية | العلاج على نفقة الدولة |
| 01/01/2009M | ايناس نظيم | رنين على الكتف الايسر | ت. صحى القاهرة |
| … | … | … | … |

Table 3: Format for each field for data

| Date | Patient Name | Scan type | Organization Name |
|---|---|---|---|
| Date | Text | Text | Text |

Three data sets A, B and C are constructed from the collected data. Data set A divides the original data into 90% for model calibration and 10% for model testing. Data set B divides the original data into 85% for model calibration and 15% for model testing. Data set C divides the original data into 80% for model calibration and 20% for model testing.

**4.2 Preprocessing**

The data collected undergoes four preprocessing steps and the data matrix is reduced from 600 rows and 4 columns, to 487 rows and 30 columns. It contains transactions for all patients in this period

The first step converts data from textual values to numeric ones in order to deal with identification numbers, table 4.

Table 4: Example of data after to converting to numeric sheet

| Data | Patient ID | Scan ID | Organization ID |
|---|---|---|---|
| 03/03/2009M | 82803 | 12 | 0 |
| 02/01/2009M | 82803 | 101 | 0 |
| 01/01/2009M | 81205 | 190 | 268 |
| 01/01/2009M | 81206 | 35 | 140 |
| 11/04/2009M | 81207 | 12 | 135 |
| 01/01/2009M | 81208 | 112 | 0 |
| 01/01/2009M | 81208 | 478 | 0 |
| 21/01/2009M | 81210 | 128 | 404 |
| … | … | … | … |

In the second step, the interesting attributes are selected which are Patient ID and Scan Type ID.

Table 5: Interesting attributes for 'RDMS' in numeric values

| Patient ID | Scan ID |
|---|---|
| 82803 | 12 |
| 82803 | 101 |
| 81205 | 190 |
| 81206 | 35 |
| 81207 | 12 |
| 81208 | 112 |

The third step converts data from numeric matrix to binary matrix, table 5. The rows of the matrix represent Patients ID while columns represent the scan type represented. Elements with value 1 indicate that the patient id did the scan type id at least once. The algorithm for converting data is shown in figure 5.

*Initialize Data matrix(number of patients, number of scan type ID)to zero;*

*Read Patient ID;*

*Read scan type ID done;*

*Data matrix (patient ID, Scan ID )=1;*

Figure 5: The algorithm for converting data to binary

The fourth step is a data filtering. It is needed as the data is a snapshot for a short period of time. Therefore, not all Patient IDs are expected to exist neither all Scan Type IDs. This is indicated in the binary matrix with either all zero row(s) or all zeros column(s), respectively. The algorithm for row elimination is in figure 6. The algorithm for column elimination is similar to that in figure 6.

*% remove Patient Id who didn't request any service (scan)*

*If there is a row whose elements = 0*

*Remove row*

*Otherwise*

*Keep it;*

Figure 6: The algorithm for filtering data

### 4.3 Clustering (K-mean)

Training sets are used to calibrate the models using WEKA software and each clustering model is then tested by the corresponding testing data. For each data set, three k-mean models are created with 10, 15 and 18 clusters; this gives a total of 9 experiments shown in table 6.

Table 6: Show experiments

| Data | 90% | 85% | 80% |
|---|---|---|---|
| | A | B | C |
| **10 Cluster** | Exp.1 | Exp.4 | Exp.7 |
| **15 Cluster** | Exp.2 | Exp.5 | Exp.8 |
| **18 Cluster** | Exp.3 | Exp.6 | Exp.9 |

### 4.4 Post processing

In this study, seven quantification levels are used to quantify the cluster centers shown in Table 7. The thirty dimension of each cluster is described by one of the seven quantification level. For example, Table 8 shows the transformation of one cluster centroid into the quantified level.

Table 7: Show the scaling of result and the evaluation of data

| Serial Number | Scaling Result | Grade Level |
|---|---|---|
| **1** | 0 | Not Done |
| **2** | 0.001 - 0.02 | Very Low |
| **3** | 0.021 - 0.04 | Low |
| **4** | 0.041 - 0.06 | Moderate |
| **5** | 0.061 - 0.08 | High-moderate |
| **6** | 0.081 - 0.09 | Very High |
| **7** | 1 | Done |

Table 8: Show the transformation of cluster into quantified

| Dimensions | Centroids value | Quantification |
|---|---|---|
| **1** | 0 | Not Done |
| **2** | 0 | Not Done |
| **3** | 0 | Not Done |
| **4** | 0 | Not Done |
| **5** | 0.021 | Low |
| **6** | 0 | Not Done |
| **7** | 0 | Not Done |
| **8** | 0 | Not Done |
| **9** | 0 | Not Done |
| **10** | 0 | Not Done |
| **11** | 0 | Not Done |
| **12** | 0.066 | High Moderate |
| **13** | 1 | Done |
| **14** | 0 | Not Done |
| **15** | 0 | Not Done |
| **16** | 1 | Done |
| **17** | 0 | Not Done |
| **18** | 0 | Not Done |
| **19** | 0 | Not Done |
| **20** | 0.081 | Very High |
| **21** | 0 | Not Done |
| **22** | 0 | Not Done |
| **23** | 0 | Not Done |
| **24** | 0 | Not Done |
| **25** | 0 | Not Done |
| **26** | 0 | Not Done |
| **27** | 0.044 | Moderate |
| **28** | 0 | Not Done |
| **29** | 0 | Not Done |
| **30** | 1 | Done |

## 5. RESULTS

The results of running the RDMS system are presented as follows:

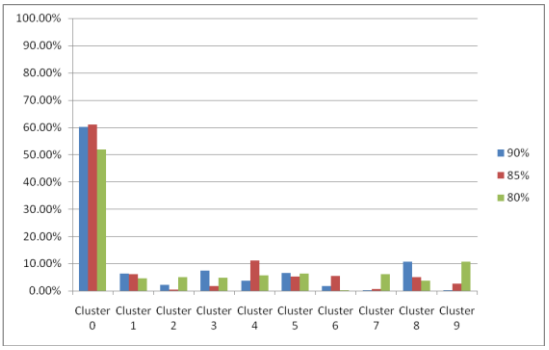


Figure 8: Distribution percentage of patients in training set for experiments A, B and C for 10 clusters model

The result is represented by the patient's distribution percent in each model for each data set. Figure 8 shows the distribution percent for each data set in case of 10 clusters model .It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

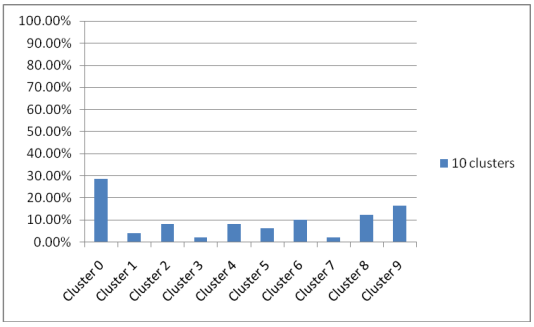


Figure 9: Distribution percentage of patients in testing set for data set B for 10 clusters model

Figure 9 describes the results of testing data. The results show that 28% of the patients exist in cluster 0; nearly 16% of the patients exist in cluster 9. It also shows that the percentage starts to decrease in the other cluster which implies that there are a lot of patients having most of their scans in clusters 0 and 9.

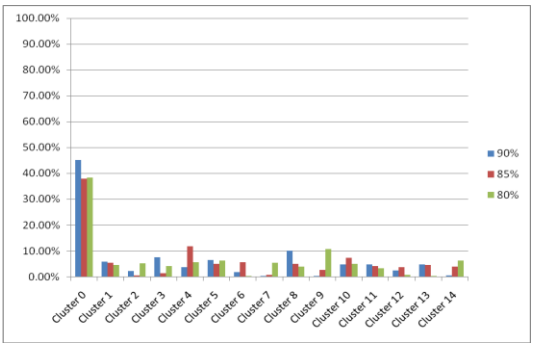


Figure 10: Distribution percentage of patients in training set for experiments A, B and C for 15 clusters model

For 15 cluster model, the result is represented by the patient's distribution percent for each data set. Figure 10 shows the distribution percent for each data set in case of 15 clusters model .It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

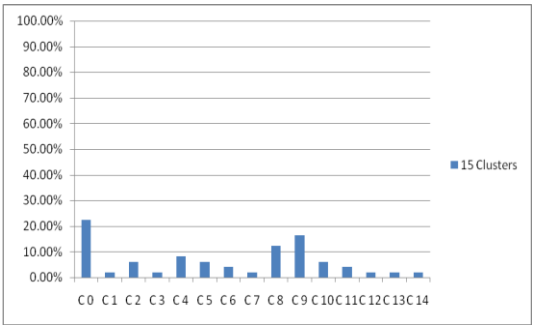


Figure 11: Distribution percentage of patients in testing set for data set B for 15 clusters model

Figure 11 describes the results of testing data. The results show that 22% of the patients exist in cluster 0; nearly 16% of the patients exist in cluster 9. It also shows that the percentage starts to decrease in the other cluster which implies that there are a lot of patients having most of their scans in clusters 0 and 9.

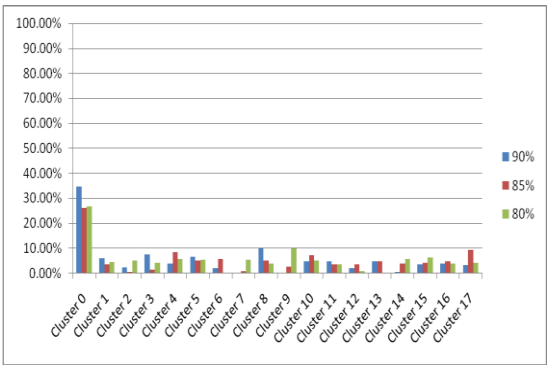


Figure 12: Distribution percentage of patients in training set for experiments A, B and C for 18 clusters model

For 18 cluster model, the result is represented by the patient's distribution percent for each data set. Figure 12 shows the distribution percent for each data set in case of 18 clusters model .It shows that the model for data set B is the most appropriate one as it describes an average between the other results.

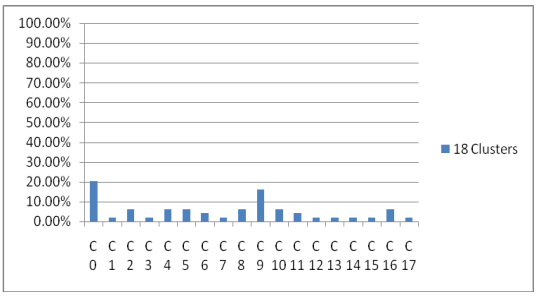


Figure 13: Distribution percentage of patients in testing set for data set B for 18 clusters model

Figure 13 describes the results of testing data. The results show that 21% of the patients exist in cluster 0; nearly 16% of the patients exist in cluster 9 almost 1% in clusters 16, 10,8,5,4 and cluster 2. It also shows that the percentage starts to decrease in the other cluster which implies that there are a lot of patients having most of their scans in clusters 0 and 9.

## 6. CONCLUSIONS

DM techniques specifically K-means succeeded in clustering patients into groups according to the requested services.

Preprocessing is an essential step in this research as it adapts medical data to computational techniques. Preprocessing step is simple and easily implemented one. Changing data formats from string or numeric to binary is essential as a requirement for K-means to work successfully. The binary matrix indicates that a patient did a specific diagnose check rather than indicating the number of times he did a specific check. This approximation has a limited effect as the data is for a short period of time.

Experimenting different K-means model are preferred in order to understand the solution for better problem understanding as well as ensure the reported results. Postprocessing helps the medical specialist to better label the different clusters and interpret the relation among clusters and their members as indicated later.

In order to fully understand the patient's clusters, the results are presented to a medical specialist to interpret them and conclude important findings.

First package "Women Scans" contains the sequence scans spatially done for women. All these scans are done every period of time for women especially in age range of (35 – 40) years such as; DEXA scan for bone density, Mammogram, and Ultra-Sound .Second package "Bone scan" contains scans such as, X-Ray for all of bone and spinal cord, MRI for all of bone and spinal cord.

Magnetic Resonance Image "MRI" scan is number of scans sometimes that ranges from low to moderate as they are scan not supported well by the center.

The most common scans done in the center are Brain, MRI Dorsal or Cervical. It is analog with a new scan that's very important for diagnoses "Ango" or "SBH" is a body haring. 3D and 4D dimensions are classified into two categories which are 3Dimensions, this scan for spatial diagnoses for the sequence of baby's growth. 4Dimensions is done to check the changes in infants and this scan is not done for diagnosis but

for checking reasons only. So it is not used most of the time.

Doppler scan, this scan is a very important scan and is not supported in most great centers because it needs high technology. This scan is classified to categories and it may be classified depending on each case of the patient for diagnose such as; Doppler for Pregnancy, Doppler for Gland, Doppler for Haring neck and ECG is very important.

Graphic brain is never done because it is not supported in this center.

After investigating the data and the results a conclusion is drawn all that generated facts nearly match with the medical fact, which means the system clusters the patients according to the requested scans. Other finding results are

- The most populated cluster is for female sector and there are special scans such as; DEKA, Ultrasound for pregnancy, Doppler for pregnancy and Ultrasound for 3&4 Diminutions.
- There are scans for older men, that can be supported by special price to get more patient such as; ECG Normal, ECG by Stress and GAMA on brain, liver and Kidney.
- There are scans used for medical checks before hiring in most organizations such as; X-RAY, MRI on Jaw and X-RAY for Bone Density.

- There are a lot of scans done because they are supported in most centers branches.

This paper is the contribution of DM in CRM for medical sector which has been rarely addressed before. RDMS is a new proposed system which is simple, straightforward with low computation needs. The proposed preprocessing component is an aggregation of several known steps. The post processing component is an optional one that eases the interpretation of the medical results. The radiology center is planning a set of actions in accordance of RDMS outcomes. The market department in the radiology center is starting to analyze the approached market sector, to introduce new scan packages and give good offers for special scan.

# References


[1]. Philip Kotler. CRM and Marketing analysis, p. 409 –p. 410, 2000.

[2]. Ibid. CRM as concepts for CRM and Customer management, p. 325, 2005.

[3]. John Johansson & Fredrik Strom. CRM, p.2-p4, 2004.

[4]. M. S. Chen, J. Han, and P. S. Yu. IEEE Trans Knowledge and Data Engineering Data mining. An overview from a database perspective, 8:866-883, 1996.

[5]. U. Fayyad, G. Piatetsky-Shapiro and W. J. Frawley. AAAI/MIT, Press definition of KDD at KDD96. Knowledge Discovery in Databases, 1991.

[6]. Gartner. Evolution of data mining, Gartner Group Advanced Technologies and Applications Research Note, 2/1/95.

[7]. International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98), 1995-1998.

[8]. R.J. Miller and Y. Yang. Association rules over interval data. SIGMOD'97, 452-461, Tucson, Arizona, 1997.

[9]. Zaki, M.J., SPADE An Efficient Algorithm for Mining Frequent Sequences Machine Learning, 42(1) 31-60, 2001.

[10]. Osmar R. Zaïane. "Principles of Knowledge Discovery in Databases - Chapter 8 Data Clustering". & Shantanu

Godbole data mining Data mining Workshop 9th November 2003.

[11]. T.Imielinski and H. Mannila. Communications of ACM. A database perspective on knowledge discovery, 39:58-64, 1996.

[12]. BIRCH Zhang, T., Ramakrishnan, R., and Livny, M. SIGMOD '96. BIRCH an efficient data clustering method for very large databases. 1996.

[13]. Pascal Poncelet, Florent Masseglia and Maguelonne Teisseire (Editors). Information Science Reference. Data Mining Patterns New Methods and Applications, ISBN 978 1599041629, October 2007.

[14]. Thearling K, Exchange Applications White Paper, Inc. increasing customer value by integrating data mining and campaign management software, 1998.

[15]. Ayman Khedr, PHD thesis. Knowledge Discovery in Databases for CRM in Egyptian public banks, p-20-p23, 2007.

[16]. Noah Gans, Spring. Service Operations Management, Vol. 5, No. 2, 2003.

[17]. Joun Mack. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. An Efficient k-Means Clustering Algorithm, Analysis and Implementation, VOL. 24, NO. 7, JULY 2002.

[18]. Andrew Moore and Brian T. Luke. Tutorial Slides, K-means and Hierarchical Clustering and K-Means Clustering, Slide 15, 2003.