Research Journal of Applied Sciences, Engineering and Technology 13(6): 427-431, 2016
DOI:10.19026/rjaset.13.3002
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 19, 2014 Accepted: February 5, 2015 Published: September 15, 2016

Research Article

A New Clustering Algorithm of Data Mining

[1]Jamal Mbarki, [1]El Miloud Jaara, [2]Sidi Yasser Eljasouli, [1]J. Mbarki and [1]E.M. Jaara
[1]Laboratory of Computer Science Research (LARI), Faculty of Sciences, Mohamed Ier University, Oujda, Morocco
[2]Integrated and Efficient Solutions, IT.sprl, Belgium

Abstract: The aim of this study is to present a new useful process of segmentation in large data, because organizing data into sensible groupings is now becoming the most fundamental modes of understanding and learning and enterprises have gathered a large amount of information over the last decades, the dilemma of managing such information by retrieving advantage in efficient way and less costing methods is becoming the key business success and takes top rows in strategy scale, different methods and techniques have been developed to reduce the data volumes to manageable structure and help enterprise to isolate the business value from the data sets. Clustering is one of those most important used data mining techniques. The algorithm that we will present can be helpful in CRM area. It can be potentially useful to better study customer profiles based on parameters called descriptor and may have a positive impact on customer retention and churn prevention, because the main aim of an ideal business is to optimise customer interactions by well remaining connected with customer.

Keywords: Clustering, CRM, customer segmentation, similarity

INTRODUCTION

One of the most important typology is based on objectives: The high level view is schematized by Fig. 1.

Classification: This method aims to assign data set to the appropriate classes: Example outcome for a credit request assignation: A bank loan representative wants to perform a customer's data analysis in order to know which customer (loan applicant) are potentially illegible or no; The underlying Model can be built as follow:

Let C = { C1, C2 } Where C1 means the Customer is illegible, C2 other way round, Ω = {X1 ... . Xn} a data set where Xi is customer data extracted from analytical financial CRM, Xij where, j belongs to the interval [1 P] represents the value of variable j for customer i, examples: Socio-demo, age, financial antecedents of a customer, second example is issued from medical ones to identify the risk factors for prostate cancer, based on clinical, diet and demographical parameters. Y**(+)** = Clinical favourable parameters, Y**(-)** for other cases.

Clustering: is an unsupervised learning task which aims to arrange the instances described in a database into a set of homogeneous and mixed clusters (Han and Kamber, 2001), where similarity intra classes and dissimilarity interclass principles are satisfied.

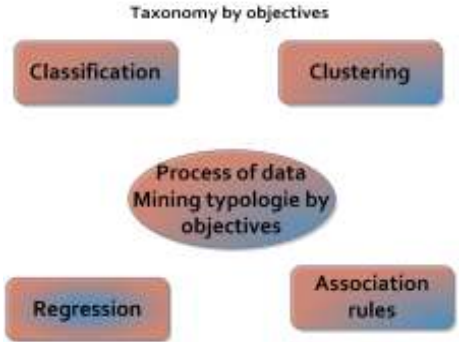

Fig. 1: Taxonomy of DM algorithms

Clustering algorithms may be grouped into two distinct types: Hierarchical and partitioning methods (Fraley and Raftery, 1998). Additional classification has been introduced later (Tan *et al*., 2002) with three new layers: Density-based methods, model-based clustering and grid based methods.

Regression: It has been a mainstay of statistics for the past 50 years and remains one of our most important tools. The principle is to predict or explain values of quantitative variable Y based on content or elements of a given variable X, X is known and true (Tibshirani, 1996), Y is called predictor, the underlying mathematical model may be described as follow (1):

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

Table 1: Data set of telco products (basket)

Transaction ID	Acquired products
775	P1, P2
875	P1, P2, P3
798	P1
890	P2 & P4
785	P1, P6
1020	P6 + P5
Support (P1)	4(66, 66%)
Support (P5)	1(16, 66%)
Support (P2)	3(50%)

$$f(x) = a_0 + \Sigma_{j=1}^{p} x_j b_j \quad (1)$$

The term a_0 is the intercept, also known as the bias in machine learning. Often it is convenient to include the constant variable 1 in X, include a_0 in the vector of coefficientsa$_0$ and then write the linear model in vector form as an inner product Eq. (2):

$$f(x) = Bx \quad (2)$$

Association rules: The aim of this task is to identify the correlated attributes and highlight eventual relationships between items of the basket, the main objective is to continuously have a focus on the customer's behaviour; Introduced by Agrawal *et al.* (1993), It is an important data mining model studied extensively by the database and data mining community. One important criterion for using such method is to suppose that all data sets are categorical. Notice some elementary notions of used algorithms such as support, least Support, Support of an item set I is the total number of tuples containing I, lest least Support Ɓ support's thresholds, frequent item set : all itemset where support >δ. Frequent item sets: the following array (Table 1) is representing a frequently purchased telco products:

If δ = 60% P1 product is frequent whereas for P5 and P2.

If we have an item set I = {A, B} we can notice the following 2 cases Fig. 2.

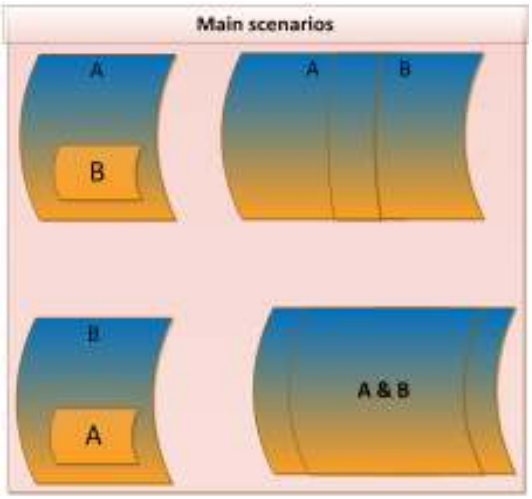

Fig. 2: Main scenarios of association rules

As recognized rule if we have two cases A, B if we have A true this is implying that we have also B true.

To asses association rules algorithm two important parameters have to be highlighted: least support and least confidante.

A rule is valid in data set if the following conditions are realized:

- $s > \delta$
- $c > \mu$

Each cluster may contain a large data set rules, rebuilding and generalizing all rules may be difficult to manage, a highly ranked representative rules may be used thanks the pruning method (Tibshirani, 1996), (Fig. 3).

For the rest of this process we will present a new algorithm of segmentation, belonging to discovery methods and that we will call M-clustering, we will summarize a methodology of its implementation in Micro Finance world.

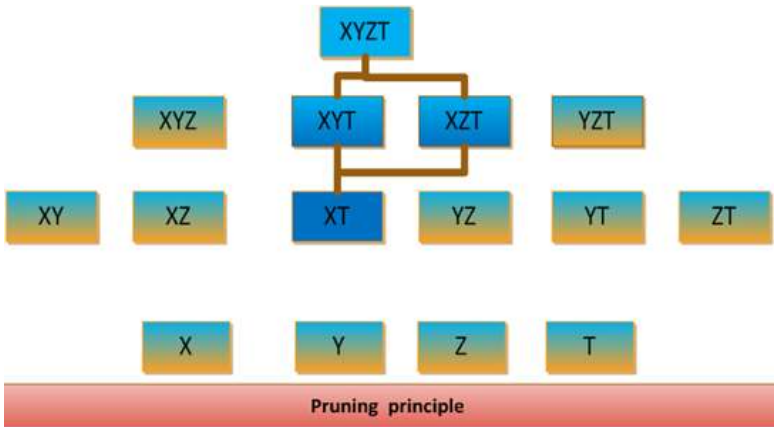

Fig. 3: Pruning method for clustering methods

BODY OF THE NEW ALGORITHM OF SEGMENTATION

For our algorithm we consider some assumptions that are mandatories for its deployment, we use Euclidean distance to calculate similarities. Let ε a data set.

- Similarity index is an application noticed S:

  $S: \varepsilon \times \varepsilon \rightarrow R^+$

  where, the following properties are satisfied:

$$\begin{cases} \forall (x,y) \in \varepsilon^2 \\ S(x,y) = S(y,x) \\ S(x,x) \geq S(x,y), S_{max} \geq S(x,y) \end{cases}$$

- Dissimilarity index is an application noticed D:

  $D: \varepsilon \times \varepsilon \rightarrow R^+$

  Where the following properties are satisfied:

$$\begin{cases} \forall (x,y) \in \varepsilon^2, where\ x \neq y \\ D(x,y) = D(y,x) \\ D(x,x) = 0 \end{cases}$$

Calculation step: We calculate similarities of each elements of the data set data ε (dimension n x p),
Let define X_0 as reference point:

$X0 = (a_0, \ldots, a_p) = 0$

$$d(x,y) = \sqrt[q]{(x_{i1} - y_{j1})^q + (x_{i2} - y_{j2})^q + \ldots\ldots (x_{ip} - y_{jp})^q}$$

$$d(xi,0) = dxi\ \sqrt[q]{(|x_{i1}|)^q + (|x_{i2}|)^q + \ldots\ldots (|x_{ip}|)^q}$$

For the rest of this article we will consider the following assumption as true:
q = 2 in this case Euclidean distances is used:

$$S = \begin{pmatrix} d(x_1,x_2) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ d(x_1,x_n) & \cdots & d(x_n,x_n) \end{pmatrix}$$

- Let K number of final desired cluster
- We calculate Max(D(Xi, Xj)) = Xk1,Xk2 ∀ Xi, Xj € Ω × Ω where Xi< > Xj;
- → We are sure that Xk1, Xk2 are belonging to 2 different clusters,
- Identify second Max(D(Xi, Xj))
  Where:
  (Xi, Xj)< > (Xk1, Xk2) = (Xk3, xk4)
  ∀ xi,yj € Ω × Ω x<>y
- → Xk1, Xk2, Xk3, Xk4 are belonging to distinct clusters.

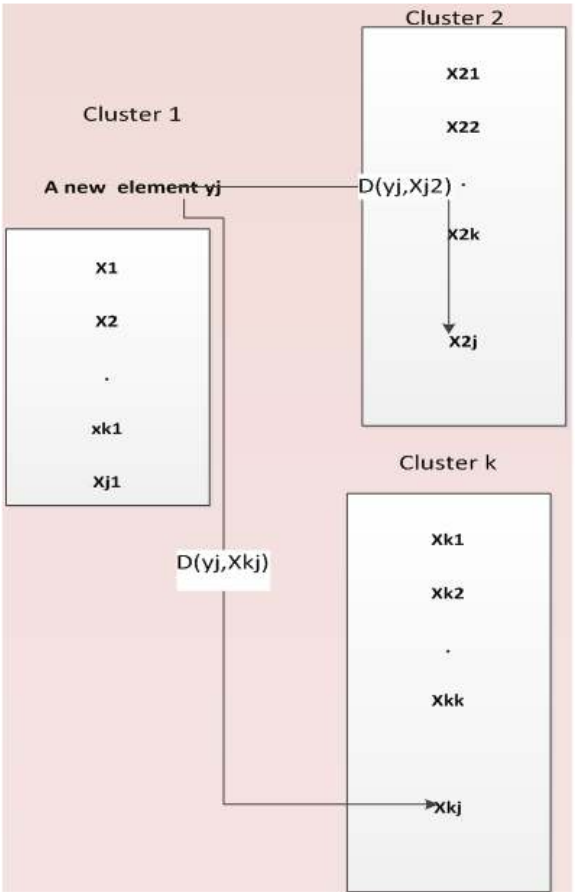

Fig. 4: High level view our algorithm steps

- Loop and (2) stop until the Xkk element is found.
- Now we have found k elements that are belonging to distinct clusters

Assignation step: This step allows to arrange all elements of the data set; it's simply assigning elements to the nearest elements of the previous step; the principle is explained as follow:

- We assign the rest of elements to the nearest K found elements
- We have now k build segments
- For each segment we perform an ascending sort of dxi.(last element is farest from the last one) (Fig. 4)

Move step: For each segment we take the last element (farest one, we assign it to nearest first element of the K-1 other segment we stop until no assignation is possible means that the segment is definitive.

MFI CUSTOMER SEGMENTATION CASE STUDY

We aim in the below case to apply this new innovating segmentation algorithm to Micro Finance

Table 2: Two sets of customer data samples

Input variable	Rational	Values
Economic data		
Household	Household size and headship	01: <25 m2; 02: 25 m2-50 m2; 03: 50 m2<
Breadwinner	Number of breadwinners	01: <2; 02: 2-3; 03: 4<
Child education	Number of children of school-going age who attend school	01: <2; 02: 2-4; 03: 4-6; 04: 6<
Household education	Education status of household	1010, Elementary; 1020; midle school; 1030, high school; 2010Undergraduate; 2020, graduate;
Asset	Land	01: No land; 02: Marginal Land <1 acre, 03: small Land (1-5 acres); large land (> 5 acres)
House comfort	Type of housing, lighting and toilet	01: Low standing; 02: Average standing; 03: Normal standing
Demographic data		
Gender	Gender	01: Male; 02: Female
Age	Age group	01: <27; 02: 27-50; 03: 50<
Marital status	Marital status	01: Single; 02: Married; 03: Divorced; 04: Widow
Disability	Disability	01: Yes; 02: No
Minority	Clients belongs to a minority	01: Yes; 02: No
Employment		01: Farming; 02 man power; 03: street selling; 04: artisan
Banking data		
Loan cycle frequency	number of previous loans	01: <2; 02: 2-3; 03: 4<
Income	monthly income	01: <15$; 02: 15$-50$; 03: 50-150$; 04<150$

Table 3: Customers segmentation after run of M-clustering

Indicator	Poor	Borderline-self sufficient	Surplus
Economic	Household size is less then 25m2, with maximum 2 breadwinners and with not or marginal land <1acre no low education.	Household size up to 50 m2, with maximum 4 breadwinners and with small land <10acre, children school going up to 4.	Household size above to 50 m2, with large land >10acre, children school going up to 4.
Demographic	67%Young, 69% of womens, divorced or widow belongs to Minorities.	57% midle age, 69% Married	71% midle age, 78% Married
Financial data	Regular casual wage labour, one-two earners	Adequate income source, small agriculture, small regular business, medium paid job has experienced one loam cycle	Assured income source, commercial agriculture, established business, high paid job

Institution (MFI) customer's. The well-known data generator http://www.mockaroo.com will be used to provide inputs data sets.

In Normal commercial banks, customer segmentation is mainly driven by Customer Relationship Management (CRM) and Profitability maximization. Banks aim and increase Customer Value Lifetime (CVL). CVL is the discounted value of the future profits that individual customers can generate.

In another hand MFI have another goal, they aims to alleviate poverty by providing financial inclusion that will assist in empowering and transferring people living in poverty to transform their lives, their children's future and their communities.

In order to achieve that mission, the Customer Social Performance Monitoring (C&SPM) is part of the key success of the positive client transformation and community up-liftment through the loans issued. It also aims at assessing the penetration of the different customer segments, which leads to another type of segmentations than the one conducted by commercial bank.

As shown in Fig. 4 this case study uses three customer data types, demographic, economic and finance details as input to the innovative M-clustering algorithm. Below we detail the used steps:

First step: We categorize the sample attribute's generated from http://www.mockaroo.com of 5200 customers, in the 3 types Demographic, Economics and Financial data see Table 1.

Second step: We will compute the similarity matrix.

Third step: Agree on the number of clusters K = 3, Then run the M-Algorithm assignation module.

Fourth step: We rollout the assignation module as much as possible till we get all the customers are assigned to one of the 3 defined clusters.

The above listed step has been experienced to two sets of customer data samples (1000 Cases and realistic one with 52000 cases) (Table 2).

M-CLUSTERING RESULTS

The M-clustering algorithm revealed the two set of data (Fig. 5), that, majority of the clients belonged to cluster2, labeled as 'borderline- Self sufficient' segment (54-50%) followed by cluster1; 'poor' segment (37-38%). Followed by cluster3: 'Suplus' customer (9-12%). This type of categorization is well known in MFI as the wealth ranking index categorization (Fig. 6 and Table 3):

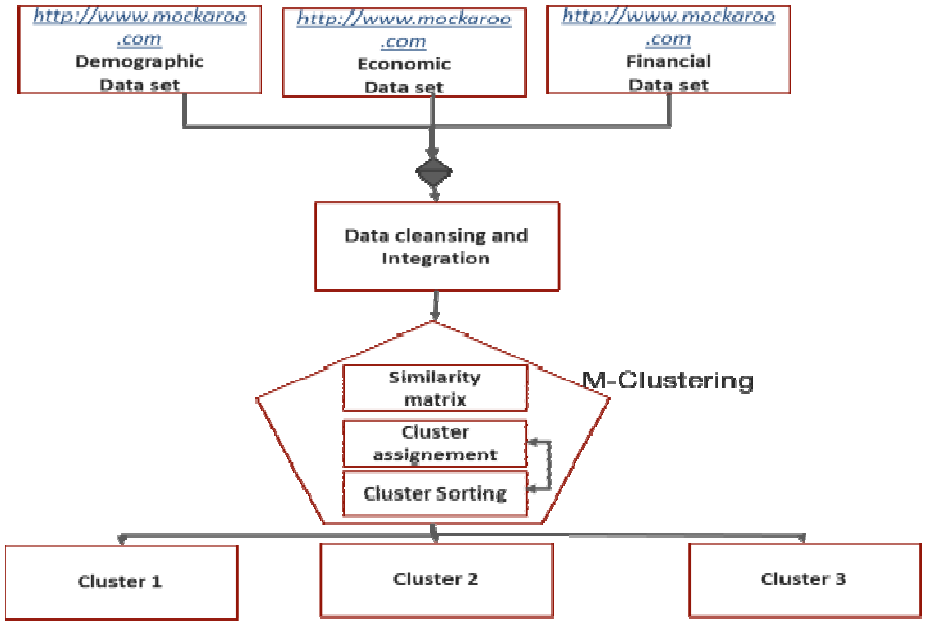

Fig. 5: Case study of M-clustering algorithm

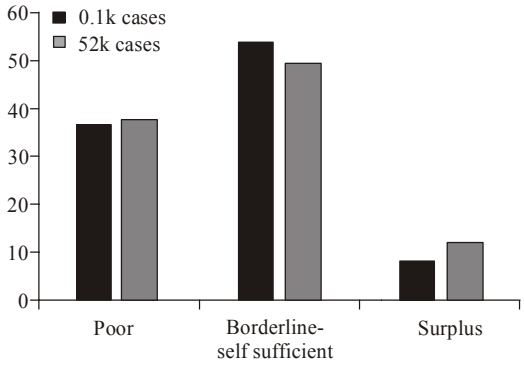

Fig. 6: Wealth ranking index representation

CONCLUSION

This study proposes a new of customer segmentation analysis using innovative M-means clustering algorithm. We have validated the algorithm to the Micro Finance Institution using the K = 3 we have proved that the algorithm can be well applied to Wealth Rank Index.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Database mining: A performance perspective. IEEE T. Knowl. Data En., 5(6): 914-925.

Fraley, C. and A.E. Raftery, 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J., 41(8): 578-588.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufman, San Francisco, Calif.

Tan, P., V. Kumar and J. Srivastava, 2002. Selecting the right interestingness measure for association patterns. Technical Report 2002-112, Army High Performance Computing Research Center.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B Met., 58(1): 267-288.