Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services

3 authors, including:

Chinedu Pascal Ezenkwu
University of Uyo
20 PUBLICATIONS   20 CITATIONS

SEE PROFILE

Simeon Ozuomba
18 PUBLICATIONS   25 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Design of New Multidisciplinary Research Network: Non-conventional Materials Science and Engineering Network View project

Project   Smart Homes for monitoring elderly people View project

Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services

Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu

Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria

Abstract—**The emergence of many business competitors has engendered severe rivalries among competing businesses in gaining new customers and retaining old ones. Due to the preceding, the need for exceptional customer services becomes pertinent, notwithstanding the size of the business. Furthermore, the ability of any business to understand each of its customers' needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation. Each segment comprises customers who share similar market characteristics. The ideas of Big data and machine learning have fuelled a terrific adoption of an automated approach to customer segmentation in preference to traditional market analyses that are often inefficient especially when the number of customers is too large. In this paper, the k-Means clustering algorithm is applied for this purpose. A MATLAB program of the k-Means algorithm was developed (available in the appendix) and the program is trained using a z-score normalised two-feature dataset of 100 training patterns acquired from a retail business. The features are the average amount of goods purchased by customer per month and the average number of customer visits per month. From the dataset, four customer clusters or segments were identified with 95% accuracy, and they were labeled: High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular-Visitors (LBIV).**

Keywords—*machine learning; data mining; big data; customer segmentation; MATLAB; k-Means algorithm; customer service; clustering; extrapolation*

I. INTRODUCTION

Over the years, the increase in competition amongst businesses and the availability of large historical data repositories have prompted the widespread applications of data mining techniques in uncovering valuable and strategic information buried in organisations' databases. Data mining is the process of extracting meaningful information from a dataset and presenting it in a human understandable format for the purpose of decision support. The data mining techniques intersect areas such as statistics, artificial intelligence, machine learning and database systems. The applications of data mining include but not limited to bioinformatics, weather forecasting, fraud detection, financial analysis and customer segmentation. The thrust of this paper is to identify customer segments in a retail business using a data mining approach. Customer segmentation is the subdivision of a business customer base into groups called customer segments such that each customer segment consists of customers who share similar market characteristics. This segmentation is based on factors that can directly or indirectly influence market or business such as products preferences or expectations, locations, behaviours and so on. The importance of customer segmentation include, inter alia, the ability of a business to customise market programs that will be suitable for each of its customer segments; business decision support in terms of risky situation such as credit relationship with its customers; identification of products associated with each segments and how to manage the forces of demand and supply; unravelling some latent dependencies and associations amongst customers, amongst products, or between customers and products which the business may not be aware of; ability to predict customer defection, and which customers are most likely to defect; and raising further market research questions as well as providing directions to finding the solutions.

Clustering has proven efficient in discovering subtle but tactical patterns or relationships buried within a repository of unlabelled datasets. This form of learning is classified under unsupervised learning. Clustering algorithms include k-Means algorithm, k-Nearest Neighbour algorithm, Self-Organising Map (SOM) and so on. These algorithms, without any knowledge of the dataset beforehand, are capable of identifying clusters therein by repeated comparisons of the input patterns until the stable clusters in the training examples are achieved based on the clustering criterion or criteria. Each cluster contains data points that have very close similarities but differ considerably from data points of other clusters. Clustering has got immense applications in pattern recognition, image analysis, bioinformatics and so on. In this paper, the k-Means clustering algorithm has been applied in customer segmentation. A MATLAB program (Appendix) of the k-Means algorithm was developed, and the training was realised using z-score normalised two-feature dataset of 100 training patterns acquired from a retail business. After several iterations, four stable clusters or customer segments were identified. The two features considered in the clustering are the average amount of goods purchased by customer per month and the average number of customer visits per month. From the dataset, four customer clusters or segments were identified and labelled thus: High-Buyers-Regular-Visitors (HBRV), High-Buyers-Irregular-Visitors (HBIV), Low-Buyers-Regular-Visitors (LBRV) and Low-Buyers-Irregular-Visitors (LBIV). Furthermore, for any input pattern that was

not in the training set, its cluster can be correctly extrapolated by normalising it and computing its similarities from the cluster centroids associated with each of the clusters. It will hence be assigned to any of clusters with which it has the closest similarity.

II. LITERATURE REVIEW

A. Customer Segmentation

Over the years, the commercial world is becoming more competitive, as such organizations have to satisfy the needs and wants of their customers, attract new customers, and hence enhance their businesses [1]. The task of identifying and satisfying the needs and wants of each customer in a business is a very complex task. This is because customers may be different in their needs, wants, demography, geography, tastes and preferences, behaviours and so on. As such, it is a wrong practice to treat all the customers equally in business. This challenge has motivated the adoption of the idea of customer segmentation or market segmentation, in which the customers are subdivided into smaller groups or segments wherein members of each segment show similar market behaviours or characteristics. According to [2], customer segmentation is a strategy of dividing the market into homogenous groups. [3] posits that "the purpose of segmentation is the concentration of marketing energy and force on subdivision (or market segment) to gain a competitive advantage within the segment. It's analogous to the military principle of concentration of force to overwhelm energy." Customer or Market segmentation includes geographic segmentation, demographic segmentation, media segmentation, price segmentation, psychographic or lifestyle segmentation, distribution segmentation and time segmentation [3].

B. Big Data

Recently, research in Big data has gained momentum. [4] defines Big data as " the word describing the large volume of both structured and unstructured data, which cannot be analyzed using traditional techniques and algorithm." According to [5], "the amount of data in our world has been exploding. Companies capture trillions of bytes of information about their customers, suppliers, and operations, and millions of networked sensors are being embedded in the physical world in devices such as mobile phones and automobiles, sensing, creating, and communicating data." Big data has demonstrated the capacity to improve predictions, save money, boost efficiency and enhance decision-making in fields as disparate as traffic control, weather forecasting, disaster prevention, finance, fraud control, business transaction, national security, education, and health care [6]. Big data is mainly characterised by three V's namely: volume, variety and velocity. There are other 2V's available - veracity and value, thus making it 5V's [4]. Volume refers to the vast amount of data in Zettabytes or Brontobytes being generated per minute; velocity refers to speed at which new data is created or the speed at which existing data moves around; variety refers to different types of data; veracity describes the degree of messiness or trustworthiness of data; and value refers to the worth of information that can be mined from data. The last V, value is what makes Big data and data mining interesting to businesses and organisations.

C. Clustering and k-Means Algorithm

According to [7], clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). [8] opined that clustering algorithms generate clusters having similarity between data objects based on some characteristics. Clustering is extensively used in many areas such as pattern recognition, computer science, medical, machine learning. [6] states that "formally cluster structure is represented as a set of subset C=C1,……..Ck of S, such that S=$\cup_{i=1}^{k}$ Ci and C_i ∩Cj=∅ for $i \neq j$. Consequently, instances in S belong to exactly one and only one subset". Clustering algorithms have been classified into hierarchical and partitional clustering algorithms. Hierarchical clustering algorithms create clusters based on some hierarchies. It is based on the idea of objects being more related to nearby objects farther away [6]. It can be top-down or bottom-up hierarchical clustering. The top-down approach is referred to as divisive while the bottom-up approach is known as agglomerative. The partitional clustering algorithms create various partitions and then evaluate them by some criterion. k-Means algorithm is one of most popular partitional clustering algorithm[4]. It is a centroid-based algorithm in which each data point is placed in exactly one of the K non-overlapping clusters selected before the algorithm is run.

The k-Means algorithm works thus: given a set of d-dimensional training input vectors { $x_1, x_2,.., x_n$ }, the *k*-Means clustering algorithm partitions the *n* training examples into *k* sets of data points or clusters $S = \{S_1, S_2, …, S_k\}$, where k≤n, such that the within cluster sum of squares is minimised.

That is,

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \epsilon s_i} \|x - \mu_i\|^2 \tag{1}$$

where, μ_i is the centroid or mean of data points in cluster s_i .

Generic k-means clustering Algorithms:

1) Decide on the number of clusters, k.

2) Initialize the k cluster centroids

3) Assign the n data points to the nearest clusters.

4) Update the centroid of each cluster using the data points therein.

5) Repeat steps 3 and 4 until the changes in positions of centroids are zero.

III. METHODOLOGY

The data used in this paper was collected from a mega retail business outfit that has many branches in Akwa Ibom state, Nigeria. The dataset consists of 2 attributes and 100 tuples, representing 100 selected customers. The two attributes include average amount of goods purchased by customer per month and average number of customer visits per month. In this paper, four steps were adopted in realising an accurate result. They include feature normalisation alongside centroids initialisation step, assignment step and updating step, which are the three major generic steps in the k-Means algorithms.

A. Feature normalisation

This is a data preparation stage. Feature normalisation helps to adjust all the data elements to a common scale in order to improve the performance of the clustering algorithm. Each data point is converted to the range of -2 to +2. Normalisation techniques include Min-max, decimal scaling and z-score. The z-score normalisation technique was used to normalise the features before running the k-Means algorithm on the dataset. Equation (2) gives the formulae for normalisation using the z-score technique.

$$x_{norm} = \frac{x - \mu_f}{\sigma_f} \qquad (2)$$

where, x_{norm} is the normalised value of x in feature vector **f**, μ_f is the meant of the feature vector **f**, and σ_f is the standard deviation of feature vector **f**.

B. Centroids Initialisation

The initial centroids or means were chosen. Figure 1 presents the initialisation of the cluster centres. Four cluster centres shown in different shapes were selected using Forgy method. In Forgy method of initialisation k (in this case k=4) data points are randomly selected as the cluster centroids.
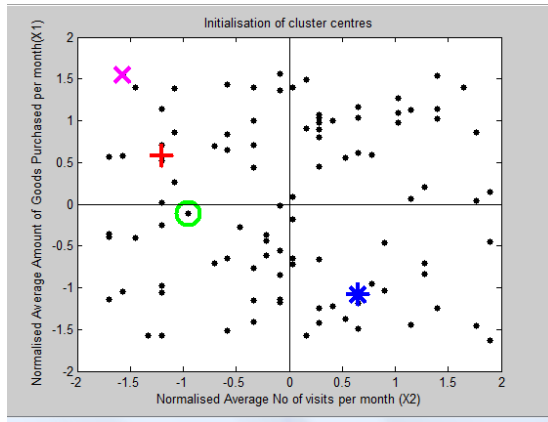

Fig. 1. The initialization stage of k-Means algorithm

C. Assignment Stage

In the assignment stage, each data point is assigned to the cluster whose centroid yields the least within cluster sum of squares compared with other clusters. That is, the square Euclidean norms of each data point from the current centroids are computed. Thereafter, the data points are assigned membership of the cluster that gives the minimum square Euclidean norm.

This has been mathematically explained in equation (3)

$$s_i^{(t)} = \{x_p : \|x_p - \mu_i^{(t)}\|^2 \le \|x_p - \mu_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\} \ (3)$$

where each data point x_p is assigned to only one cluster or set $s^{(t)}$ at the iteration t.

D. Updating Stage

After each iteration, new centroid is computed for each cluster as the mean of all the data points present in the cluster as shown in equation (4)

$$\mu_t^{(t+1)} = \frac{1}{|s_i^{(t)}|} \sum_{x_j \in s_i^{(t)}} x_i \qquad (4)$$

where, $\mu_t^{(t+1)}$ is the updated centroid.

Fig. 2 presents the positions of the centroids and the updated assignment of their cluster members after the 30th iteration. The each cluster members assume the same shapes as their cluster centroid. Table II shows the changes in the cluster centroids from the initialisation stage (0th iteration) to the 5th iteration.
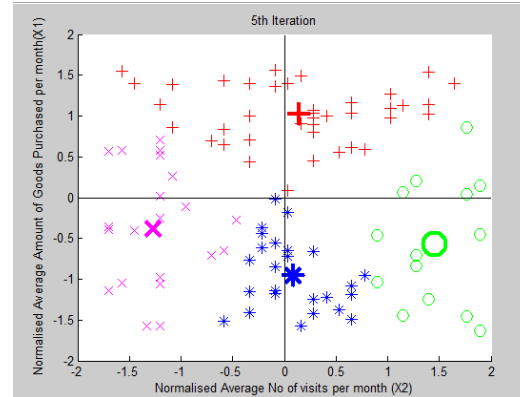

Fig. 2. Positions of the centroids and their cluster members after the 30th iteration

TABLE I.   INITIALISATION AND UPDATING OF THE CLUSTER VECTORS OR CENTROIDS)

Iteration	INTIALISED CLUSTER CENTROIDS:			
	Cluster Centre +	Cluster Centre *	Cluster Centre O	Cluster Centre X
0	-0.0892 1.3654	0.6541 -1.0856	-0.2131 -0.3669	-0.2131 -0.3669
	UPDATED CLUSTER CENTROIDS:			
1	0.5656 1.0971	0.8733 -0.9508	-0.6306 -0.6728	-0.6306 -0.6728
2	0.5798 1.0456	0.9976 -0.9639	-0.5466 -0.8295	-0.5466 -0.8295
3	0.5502 1.0346	1.0376 -0.9348	-0.5600 -0.9284	-0.5600 -0.9284
4	0.5502 1.0346	1.0376 -0.9348	-0.5641 -0.9557	-0.5641 -0.9557
5	0.5502 1.0346	1.0376 -0.9348	-0.5901 -0.9894	-0.5901 -0.9894

IV.   RESULTS AND DISCUSSION

The k-Means clustering algorithm converged after 100 iterations. That is, the cluster centroids became stable. Figure 3 shows the graph of the converged data points and centroids. After this, the k-Means algorithm was able to cluster almost the entire data points correctly. The centroids or the cluster vectors after convergence are:

*Cluster Centre +   Cluster Centre *   Cluster Centre O   Cluster Centre X*
[-0.8325   0.9574] [0.7403   -1.0926]   [-0.8279   -0.7217]   [0.8444   0.8412]

Each of the clusters represents a customer segment. From Figure 3, the data points at the right hand top corner represent HBRV; the data points left hand top corner represent the HBIV; the data points at the right hand lower corner represent LBRV; while those at the left hand lower corner represent the LBIV. This is clearly shown in Table II.

TABLE II.   DESCRIPTION OF EACH CLUSTER IN TERMS OF THE CUSTOMER
SEGMENT

HBIV	HBRV
Cluster +	*Cluster* **X**
LBIV	LBRV
Cluster **O**	*Cluster* *

V.   PERFORMANCE EVALUATION

Purity measure was used to measure the extent to which a cluster contains of class of data points. The purity of each cluster is computed with equation (5).

$$purity(D_i) = \max_j \left(P_i(C_j) \right) \tag{5}$$

Where, $P_i(C_j)$ is the proportion of class C_j data points in cluster i or D_i.

The total purity of the whole clustering i.e. considering all the clusters is given by equation (6).

$$Purity_{total}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} X \ purity(D_i) \tag{6}$$

Where, D is the total number of data points being classified.

The confusion matrix is presented in Table III.

TABLE III.   CONFUSION MATRIX

Cluster	HBIV	HBRV	LBIV	LBRV	Purity
Cluster +	21	1	0	0	0.954
Cluster X	0	28	0	0	1.000
Cluster O	2	0	24	1	0.889
Cluster *	0	0	1	22	0.957
Total	23	29	25	23	0.950

Since, $Purity_{total}(D)$ = 0.95(from row 6, column 6 of Table 3), the clustering algorithm was 95% accurate in performing the customers segmentation.

VI.   CONCLUSIONS

This paper has presented a MATLAB implementation of the k-Means clustering algorithm for customer segmentation based on data collected from a mega business retail outfit that has many branches in Akwa Ibom state, Nigeria. The algorithm has a purity measure of 0.95 indicating 95% accurate segmentation of the customers. Insight into the business's customer segmentation will avail it with the following advantages: the ability of the business to customise market programs that will be suitable for each of its customer segments; business decision support in terms of risky situations such as credit relationship with its customers; identification of products associated with each segments and how to manage the forces of demand and supply; unravelling some latent dependencies and associations amongst customers, amongst products, or between customers and products which the business may not be aware of; ability to predict customer defection and which customers are most likely to defect; and raising further market research questions as well as providing directions to finding the solutions.
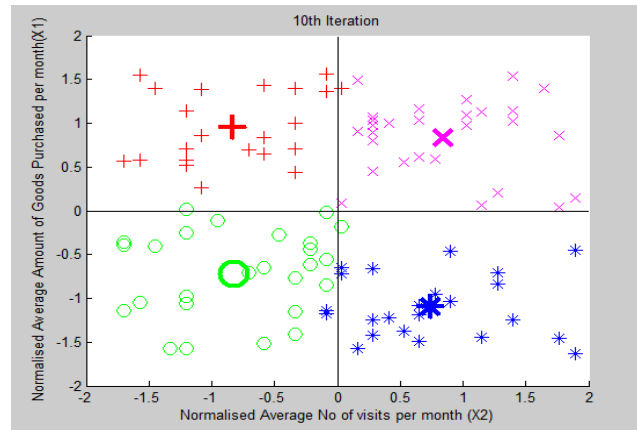

Fig. 3.   The centroids converge after 100th iteration

REFERENCES

[1] Puwanenthiren Premkanth, "Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC." Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1.

[2] Sulekha Goyat. "The basis of market segmentation: a critical review of literature". European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online).Vol 3, No.9, 2011

[3] *By Jerry W Thomas."*Market Segmentation". 2007. Retrieved from www.decisionanalyst.com on 12-July, 2015.

[4] T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. "Survey on mining clusters using new k-mean algorithm from structured and unstructured data". International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2.

[5] McKinsey Global Institute. Big data. The next frontier for innovation, competition, and productivity. 2011. Retrieved from www.mckinsey.com/mgi on 14 July, 2015.

[6] Jean Yan. "Big Data, Bigger Opportunities- Data.gov's roles: Promote, lead, contribute, and collaborate in the era of big data". 2013. Retrieved from http://www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf on 14 July 2015.

[7] A.K. Jain, M.N. Murty and P.J. Flynn."Data Clustering: A Review". ACM Computing Surveys. 1999. Vol. 31, No. 3.

[8] Vaishali R. Patel1 and Rupa G. Mehta. "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm". IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011 ISSN (Online): 1694-0814

[9] Md. Al-Amin Bhuiyan and Hiromitsu Hama, "Identification of Actors Drawn in Ukiyoe Pictures", Pattern Recognition, Vol. 35, No. 1, pp. 93-102, 2002.

[10] S. O. Olatunji, M. Al-Ahmadi, M. Elshafei, and Y. A. Fallatah, "Saudi arabia stock prices forecasting using artificial neural networks," pp.81–86, 2011.

[11] Q. Wen, Z. Yang, Y. Song, and P. Jia, "Automatic stock decision support system based on box theory and svm algorithm," *Expert System Application*, vol. 37, no.2, pp. 1015–1022, Mar. 2010.[Online]. Available: http://dx.doi.org/10.1016/j.eswa.2009.05.093.

[12] P.-C. Chang, C.-Y. Fan, and J.-L. Lin, "Trend discovery in financial time series data using a case based fuzzy decision tree," *ExpertSystem Application*, vol. 38, no. 5, pp. 60706080, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2010.11.006

APPENDIX

```
clc;clf;close;clear all;
load CustData % Data file containing 100-by-2 training examples, X
%Normalisation and Selection of initial centroids
```

```
X=[(X(:,1)-mean(X(:,1)))/std(X(:,1))     (X(:,2)-
mean(X(:,2)))/std(X(:,2))];
j = 1;k=1;l=1;
i = randi(length(X));
while j==i
    j=randi(length(X));
end
while k==i|k==j
    k =randi(length(X));
end
while l==i|l==j|l==k;
    l =randi(length(X));
end
centr1 = X(i,:);centr2 = X(j,:); centr3 = X(k,:);centr4 = X(l,:);
%Initial plots of points and position of initial centroids
plot(X(:,1),X(:,2),'.k','MarkerSize',15)
hold on
plot(centr1(1),centr1(2),'+r','MarkerSize',18,'LineWidth',3)
plot(centr2(1),centr2(2),'*b','MarkerSize',18,'LineWidth',3)
plot(centr3(1),centr3(2),'Og','MarkerSize',18,'LineWidth',3)
plot(centr4(1),centr4(2),'Xm','MarkerSize',18,'LineWidth',3)
title('Initialisation of cluster centres')
xlabel('Normalised Average No of visits per month (X2)')
ylabel('Normalised Average Amount of Goods Purchased per
month(X1)')
hold off;
%Iterations to update Centroids and assign clusters members
count = 1;
while count <=10
    d1=(X-[ones(length(X),1)*centr1(1)
            ones(length(X),1)*centr1(2)]).^2;
    d2=(X-[ones(length(X),1)*centr2(1)
            ones(length(X),1)*centr2(2)]).^2;
    d3=(X-[ones(length(X),1)*centr3(1)
            ones(length(X),1)*centr3(2)]).^2;
    d4=(X-[ones(length(X),1)*centr4(1)
            ones(length(X),1)*centr4(2)]).^2;
    d11 = d1(:,1)+d1(:,2);
    d22 = d2(:,1)+d2(:,2);
    d33 = d3(:,1)+d3(:,2);
    d44 = d4(:,1)+d4(:,2);
    row1 = d11<d22 & d11<d33 & d11<d44;
    row2 = d22<d11 & d22<d33 & d22<d44;
    row3 = d33<d22 & d33<d11 & d33<d44;
    row4 = d44<d22 & d44<d11 & d44<33;
    cluster1 = X(row1,:);
    cluster2 = X(row2,:);
    cluster3 = X(row3,:);
    cluster4 = X(row4,:);
    centr1 = [mean(cluster1(:,1)) mean(cluster1(:,2))];
    centr2 = [mean(cluster2(:,1)) mean(cluster2(:,2))];
    centr3 = [mean(cluster3(:,1)) mean(cluster3(:,2))];
    centr4 = [mean(cluster4(:,1)) mean(cluster4(:,2))];
    count = count + 1;
end
% Plot the final centroids positions and cluster data points
figure; hold on;
plot(cluster1(:,1),cluster1(:,2),'+r','MarkerSize',10)
plot(cluster2(:,1),cluster2(:,2),'*b','MarkerSize',10)
plot(cluster3(:,1),cluster3(:,2),'og','MarkerSize',10)
plot(cluster4(:,1),cluster4(:,2),'Xm','MarkerSize',10)
plot(centr1(1),centr1(2),'+r','MarkerSize',18,'LineWidth',3)
plot(centr2(1),centr2(2),'*b','MarkerSize',18,'LineWidth',3)
plot(centr3(1),centr3(2),'Og','MarkerSize',18,'LineWidth',3)
plot(centr4(1),centr4(2),'Xm','MarkerSize',18,'LineWidth',3)
plot([-2 0 2],[0 0 0],'-k')
plot([0 0 0],[-2 0 2],'-k')
title('100th Iteration')
xlabel('Normalised Average No of visits per month (X2)')
ylabel('Normalised Average Amount of Goods Purchased per
month(X1)')
```