ỨNG DỤNG THUẬT TOÁN K-MEANS VÀO BÀI TOÁN PHÂN NHÓM KHÁCH HÀNG

APPLICATION OF K-MEANS ALGORITHM TO CUSTOMER SEGMENTATION PROBLEM

Ngày nhận bài : 15/9/2020
Ngày nhận kết quả phản biện : 14/12/2020
Ngày duyệt đăng : 15/12/2020

Nguyễn Thị Ngọc Hạnh - Phạm Ngọc Lâm
Trường Đại học Tài chính - Kế toán

TÓM TẮT

Phân nhóm khách hàng là cách phân loại và sắp xếp khách hàng vào từng nhóm dựa trên các đặc điểm chung như vị trí địa lí, nhu cầu, hành vi của khách hàng… nhằm giúp doanh nghiệp xác định chiến lược marketing phù hợp với từng nhóm khách hàng. Trong bài báo này, chúng tôi sử dụng thuật toán K-Means để phân nhóm khách hàng dựa vào bộ dữ liệu khách hàng mà doanh nghiệp đã thu thập được. Đây là một trong những thuật toán phân cụm đơn giản nhưng hiệu quả, và có thể thực hiện trên bộ dữ liệu lớn.

Từ khóa: *Thuật toán K-Means, phân nhóm khách hàng.*

ABSTRACT

Customer segmentation is a way of classifying and arranging customers into groups based on common features such as geographical locations, customer needs, customer behaviors, etc. to help businesses determine marketing strategies for each group appropriately. In this article, we use K-Means algorithm to group customers based on the customer data that the business has collected. This algorithm is one of the simple clustering algorithms, but it is effective, and it can be done on big data.

Key words: *K-Means algorithm, customer segmentation*

1. Giới thiệu

Cùng với sự hội nhập và phát triển hiện nay của nền kinh tế thị trường, doanh nghiệp phải cạnh tranh với nhiều nhà cung cấp với các sản phẩm đa dạng, chất lượng khác nhau. Vì vậy, khách hàng ít trung thành với một nhà cung cấp mà có thể chuyển từ nhà cung cấp này sang nhà cung cấp khác để có giá, sản phẩm và dịch vụ tốt nhất. Chính vì thế việc quản lý và chăm sóc khách hàng sẽ giúp doanh nghiệp xây dựng được mối quan hệ lâu dài với khách hàng. Tuy nhiên, khi số lượng khách hàng tăng lên quá nhiều, doanh nghiệp sẽ gặp rất nhiều khó khăn trong việc kiểm soát dịch vụ chăm sóc khách hàng để đảm bảo khách hàng luôn cảm thấy hài lòng với dịch vụ, sản phẩm của doanh nghiệp. Do đó, phân nhóm khách hàng là bước quan trọng mà doanh nghiệp cần làm lúc này.

Có nhiều cách để phân nhóm khách hàng. Trước đây, bộ phận marketing phân nhóm chủ yếu dựa vào các thông tin truyền thống như vị trí địa lý, giới tính, tuổi tác, dân tộc, thu nhập, học vấn… Ngày nay, với sự phát triển của khoa học dữ liệu, các doanh nghiệp luôn chú trọng đến việc lưu trữ cẩn thận các cơ sở dữ liệu khách hàng như hóa đơn khách hàng, chi tiết hóa đơn…Từ cơ sở dữ liệu đó, họ tìm cách sắp xếp các khách hàng vào các nhóm đối tượng khác nhau để có thể áp dụng những chiến lược kinh doanh cụ thể cho từng nhóm đối tượng. Điều này giúp cho khách hàng được tiếp cận với các sản phẩm thật sự phù hợp với bản thân họ. Sự phù hợp đó sẽ kéo doanh số của doanh nghiệp tăng lên.

Vấn đề đặt ra là làm sao có thể phân nhóm khách hàng khi mà số lượng hóa đơn rất lớn. Kỹ thuật phân cụm trong khai phá dữ liệu có thể giải quyết được vấn đề này. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm, sao cho các đối tượng trong cùng một cụm tương tự nhau và các đối

tượng khác cụm thì không tương tự nhau. Kỹ thuật phân cụm đã được sử dụng trong nhiều lĩnh vực khác nhau như marketing, sinh học, y tế, địa lý, khai phá web... Hiện nay có rất nhiều thuật toán để phân cụm dữ liệu bao gồm phân cụm phân cấp Hierarchical Clustering, phân cụm theo mật độ DBSCAN, phân cụm mô hình EM... Trong bài viết này chúng tôi chọn thuật toán phân cụm đơn giản và phổ biến đó là thuật toán K-Means.

Cấu trúc bài báo gồm các phần: Giới thiệu, thuật toán K-Means, mô tả dữ liệu, áp dụng thuật toán K-Means để phân nhóm khách hàng, kết luận.

2. Thuật toán K-Means

Thuật toán K-Means được đề xuất bởi MacQueen là một trong những thuật toán thông dụng nhất trong phân nhóm dữ liệu. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (điểm) đã cho vào k cụm (k là số các cụm được xác định trước, k nguyên dương) sao cho tổng bình phương khoảng cách giữa các điểm đến tâm cụm là nhỏ nhất.

Thuật toán K-Means thực hiện qua các bước chính sau:[1]

Bước 1: Chỉ định số lượng cụm k

Bước 2: Chọn ngẫu nhiên k điểm từ tập dữ liệu làm tâm cho k cụm

Bước 3: Tính khoảng cách giữa các đối tượng đến k tâm (thường dùng khoảng cách Euclidean)

Bước 4: Nhóm các đối tượng vào nhóm gần nhất

Bước 5: Xác định lại tâm mới cho các nhóm bằng cách tính giá trị trung bình cho các điểm dữ liệu trong các cụm tương ứng.

Bước 6: Thực hiện lại bước 3 cho đến khi không có sự thay đổi nhóm nào của các điểm dữ liệu

3. Mô tả dữ liệu

Dữ liệu được thu thập từ một nhà phân phối của Công ty Masan tại địa bàn tỉnh Quảng Nam, với các sản phẩm đa dạng về năm nhóm mặt hàng bao gồm: Gia vị (Dầu hào, Hạt nêm, Chinsu, Tam thái tử...), Mì (Kokomi, Omachi, Tiến vua...), Cà phê (Café phil 2 in 1, Vinacafe 3 in 1, Café wake up Sài Gòn...), Ngũ cốc (Ngũ Cốc B'fast, Ngũ cốc VCF) và Thịt (Thịt viên, Xúc xích cao bồi, Xúc xích Ponnie...). Bộ dữ liệu gồm 580 hóa đơn bán hàng cho các cửa hàng (khách hàng) trên địa bàn tỉnh Quảng Nam trong tháng 12/2019. Sau khi thu thập chúng tôi đã thực hiện các thao tác tiền xử lí để được bảng dữ liệu tổng hợp gồm 580 bản ghi, mỗi bản ghi có 6 thuộc tính. Bảng dữ liệu được lưu trữ trong file Data_Kmeans.csv.

Chúng tôi sử dụng ngôn ngữ lập trình Python để xử lí file dữ liệu và phân nhóm 580 khách hàng này.

	ID_KH	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	123977	930,000	452,000	630,000	0	194,400
1	123987	12,217,800	843,000	3,224,000	0	0
2	123988	12,043,000	4,561,000	1,706,800	364,000	262,400
3	123989	26,080,000	5,380,000	3,079,000	637,000	582,400
4	123991	30,413,000	11,137,430	8,368,570	0	2,343,600
..	...	...	...	...	...	...
575	2474124	415,800	0	0	0	0
576	2474125	0	761,000	665,000	0	388,800
577	2474126	56,630,800	0	0	0	0
578	2474127	1,364,500	0	0	0	0
579	2474128	2,278,800	0	1,572,000	0	0

Hình 1. Bảng dữ liệu

Tiến hành lập bảng thống kê mô tả bộ dữ liệu để biết thêm một số thông tin về bộ dữ liệu đã thu thập được. Việc lập bảng thống kê được thực hiện đơn giản bằng cách gọi hàm describe() trong Python.

	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
count	580	580	580	580	580
mean	4,921,488.71	825,080.58	1,129,646.85	41,731.19	65,083.36
std	12,699,412.19	2,035,379.91	5,799,543.10	198,699.09	239,751.15
min	0.00	0.00	0.00	0.00	0.00
25%	478,350.00	0.00	0.00	0.00	0.00
50%	1,245,000.00	166,000.00	0.00	0.00	0.00
75%	5,168,400.00	687,000.00	130,250.00	0.00	0.00
max	164,268,000.00	27,690,000.03	63,495,578.51	3,154,421.49	2,788,800.00

Hình 2. Bảng thống kê dữ liệu

4. Áp dụng thuật toán K-Means để phân nhóm khách hàng

Trước khi áp dụng thuật toán K-Means trong ngôn ngữ lập trình Python để phân cụm khách hàng thành k nhóm riêng biệt dựa vào tập dữ liệu trên, ta phải chuẩn hóa dữ liệu bằng cách sử dụng chức năng MinMaxScalar trong Python.

Với tập dữ liệu thu thập được, sử dụng thuật toán Elbow Method trong [7] ta xác định được số cụm tối ưu là 4. Thực hiện thuật toán K-Means trên bộ dữ liệu sau khi đã chuẩn hóa với số cụm được chỉ định là k= 4, ta thu được kết quả 4 nhóm khách hàng riêng biệt.

	ID_KH	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	123988	12,043,000	4,561,000	1,706,800	364,000	262,400
1	123989	26,080,000	5,380,000	3,079,000	637,000	582,400
2	124008	40,212,000	16,032,000	368,000	0	0
3	124013	26,382,000	8,229,494	8,915,306	0	0
4	124014	5,913,000	1,881,000	3,049,000	637,000	453,200
..	...	...	...	...	...	...
40	2148590	15,594,200	5,774,000	0	0	97,200
41	2148596	22,594,000	4,632,000	0	637,000	0
42	2235504	4,013,300	3,870,740	3,794,413	263,847	388,000
43	2443500	11,067,000	1,868,559	16,366,441	0	129,600
44	2474126	56,630,800	0	0	0	0

Hình 3. Dữ liệu của nhóm 0

	ID_KH	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	124027	5,766,600	1,296,917	46,326,083	0	194,000
1	124079	164,268,000	9,351,000	51,072,000	1,274,000	0
2	124779	46,192,800	5,791,000	46,923,000	0	0
3	134272	582,000	0	33,038,000	637,000	0
4	2082577	154,828,000	13,768,000	63,495,579	3,154,421	0
5	2290727	0	522,000	41,483,000	0	194,400
6	2290735	0	0	40,131,000	637,000	0
7	2290737	0	0	46,237,000	0	0

Hình 4. Dữ liệu của nhóm 1

	ID_KH	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	123977	930,000	452,000	630,000	0	194,400
1	123987	12,217,800	843,000	3,224,000	0	0
2	123992	5,341,800	344,000	0	0	0
3	123995	5,837,400	166,000	0	0	0
4	124009	5,043,000	0	0	0	0
..	...	...	...	...	...	...
513	2474123	239,400	166,000	0	0	0
514	2474124	415,800	0	0	0	0
515	2474125	0	761,000	665,000	0	388,800
516	2474127	1,364,500	0	0	0	0
517	2474128	2,278,800	0	1,572,000	0	0

Hình 5. Dữ liệu của nhóm 2

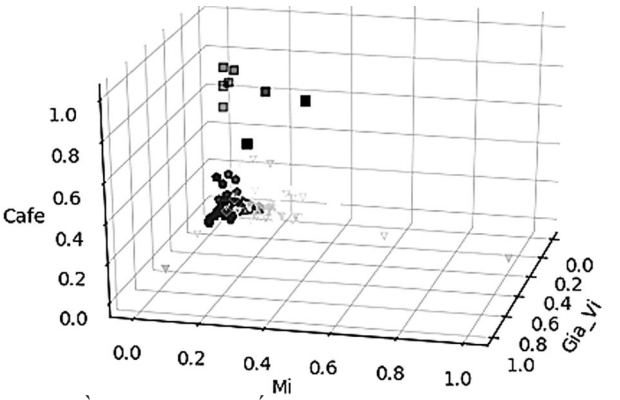
	ID_KH	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	123991	30,413,000	11,137,430	8,368,570	0	2,343,600
1	124015	16,744,000	5,041,009	20,371,275	987,716	1,164,000
2	124024	6,532,500	1,141,500	3,160,000	0	1,164,000
3	124792	11,246,800	6,188,000	8,359,000	364,000	1,219,200
4	134685	22,118,000	5,822,000	630,000	637,000	2,282,000
5	2082068	5,348,400	4,640,000	1,117,200	630,000	2,788,800
6	2147040	1,356,000	4,822,000	3,936,000	360,000	1,672,800
7	2228058	12,811,000	676,000	0	728,000	946,000
8	2428608	6,142,800	1,930,000	205,000	0	1,219,200

Hình 6. Dữ liệu của nhóm 3

Để có cái nhìn tổng quan về các nhóm, chúng tôi tiến hành thống kê dữ liệu trung bình của từng mặt hàng theo từng nhóm.

	cluster	Gia_Vi	Mi	Cafe	Ngu_Coc	Thit
0	Nhom_0	20,299,310.04	5,008,850.00	2,924,779.26	222,757.41	209,612.18
1	Nhom_1	46,454,675.00	3,841,114.61	46,088,207.71	712,802.69	48,550.00
2	Nhom_2	2,812,055.21	349,462.60	209,898.46	9,210.18	25,343.24
3	Nhom_3	12,523,611.11	4,599,771.02	5,127,449.43	411,857.33	1,644,400.00

Hình 7. Bảng thống kê dữ liệu trung bình theo nhóm



Hình 8. Đồ thị minh họa kết quả phân nhóm trên 3 thuộc tính

Chúng ta xem xét số liệu từng nhóm, kết hợp với các bảng thống kê để tìm ra một số đặc điểm chung của mỗi nhóm.

Nhóm 0 (Hình 3) gồm 45 khách hàng, đa số các khách hàng thuộc nhóm này đều mua cả gia vị (Gia_Vi) và mì (Mi) với số tiền khá cao, nhiều khách hàng mua đầy đủ 5 mặt hàng. Số tiền trung bình của mỗi mặt hàng thuộc nhóm này khá cao so với mức trung bình của toàn dữ liệu.

Nhóm 1 (Hình 4) có 8 khách hàng, các khách hàng này chủ yếu mua mặt hàng cà phê (Cafe) với số tiền rất lớn nằm trong khoảng từ 33.038.000 đồng đến 63.495.579 đồng. Trong khi số tiền trung bình của tất cả các khách hàng mua mặt hàng này chỉ là 1.129.646 đồng. Khách hàng thuộc nhóm này chỉ có 2 khách hàng mua thịt (Thit) và có 2 khách hàng mua gia vị với số tiền rất lớn đó là khách hàng số 1 và số 4 trong nhóm 1.

Nhóm 2 (Hình 5) là nhóm có nhiều khách hàng nhất gồm 518 khách hàng, đa số các khách hàng thuộc nhóm này đều mua cả gia vị và mì nhưng với số tiền ở mức thấp, dưới trung bình. Hầu như các khách hàng thuộc nhóm này không mua ngũ cốc (Ngu_Coc) và thịt. Số tiền trung bình của mỗi mặt hàng thuộc nhóm này thấp so với mức trung bình của toàn dữ liệu.

Nhóm 3 (Hình 6) có 9 khách hàng, các khách hàng trong nhóm này đều mua mặt hàng thịt với số tiền lớn trong khoảng từ 946.000 đồng đến 2.788 .800 đồng. Trong khi số tiền trung bình của tất

cả các khách hàng mua mặt hàng này chỉ là 65.083.000 đồng. Khách hàng thuộc nhóm này gần như mua đầy đủ cả 5 mặt hàng.

5. Kết luận

Bài viết này sử dụng thuật toán K-Means trong ngôn ngữ lập trình Python để thực hiện phân nhóm 580 khách hàng của nhà phân phối Masan tại địa bàn tỉnh Quảng Nam, dựa vào dữ liệu tháng 12/2019. Kết quả đã thu được 4 nhóm khách hàng với các đặc trưng riêng biệt.

Việc phân nhóm khách hàng dựa vào tập dữ liệu ban đầu có ý nghĩa rất lớn đối với các doanh nghiệp. Sau khi thu được các nhóm, bộ phận marketing sẽ tìm hiểu, rút ra thông tin chung của từng nhóm. Khi đã nắm bắt được hành vi mua sắm của từng nhóm khách hàng, doanh nghiệp có thể đưa ra các chiến lược quảng cáo, khuyến mãi riêng cho từng nhóm khách hàng. Những chiến lược đó sẽ làm hài lòng từng đối tượng khách hàng và kéo theo doanh thu của doanh nghiệp sẽ tăng lên.

TÀI LIỆU THAM KHẢO

1. Lê Hồng Diễn, Nguyễn Phúc Sơn, Phạm Hoàng Uyên, Lê Văn Hinh (2019), Bài toán phân nhóm đối với khách hàng mua sắm tại siêu thị Coopextra Thủ Đức, Tạp chí Phát triển Khoa học và Công nghệ – Kinh tế - Luật và Quản lý, 3(1):28- 36.
2. Nguyễn Văn Lễ, Mạnh Thiên Lý, Nguyễn Thị Định, Nguyễn Thị Thanh Thủy (2018), Cải tiến thuật toán k-means và ứng dụng hỗ trợ sinh viên chọn chuyên ngành theo học chế tín chỉ, Tạp chí Khoa học công nghệ và Thực phẩm 15 (1) (2018) 152-160.
3. Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu (2015), Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services, International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.10, 2015.
4. Fahim, A.M., Salem A.M., Torkey F.A., Ramadan M.A. (2006), An efficient enhanced k-means clustering algorithm, J Zhejiang Univ SCIENCE A 2006 7(10):1626-1633.
5. Kishana R. Kashwan, Member, IACSIT, and C. M. Velu (2013), Customer Segmentation Using Clustering and Data Mining Techniques, International Journal of Computer Theory and Engineering, Vol. 5, No. 6, December 2013.
6. L. Ye, C. Qiu-ru, X. Hai-xu, L. Yi-jun and Y. Zhi-min (2012), Telecom customer segmentation with K-means clustering, 2012 7th International Conference on Computer Science & Education (ICCSE), Melbourne, VIC, 2012, pp. 648-651.
7. Purnima Bholowalia, Arvind Kumar (2014), EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN, International Journal of Computer Applications (0975 – 8887)Volume 105 – No. 9, November 2014.
8. Yakur, M. A., et al (2018), Integration K-Means clustering method and elbow method for identification of the best customer profile cluster, IOP Conference Series: Materials Science and Engineering, Vol. 336. No. 1. IOP Publishing, 2018.