Int. J. Nonlinear Anal. Appl.
Volume 12, Special Issue, Winter and Spring 2021, 633-642
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.22324.2350

Customer Segmentation for Life Insurance in Iran Using K-means Clustering

Farzan Khamesian[a], Farbod Khanizadeh[a,*], Alireza Bahiraie[b]

[a]*Insurance Research Center (IRC), Tehran, Iran*
[b]*Department of Mathematics, Faculty of Mathematics, Statistics & Computer Science, Semnan University, Iran*

(Communicated by Madjid Eshaghi Gordji)

Abstract

Concerning life insurance, penetration rate is one of the main goal of every developed insurance industry. In this sense systematic marketing is a significant component in strategic plan of insurance companies. To achieve the goal insurers need to group their client into different groups in which some common features are shared and people demonstrate a similar pattern. This paper utilizes K-means clustering as an unsupervised learning algorithm in order to divide customers into number of clusters. The clusters are constructed based on two independent variables namely; car and life insurance premiums. Then the descriptive statistics of other determining features are provided with which the most willing group in purchasing life insurance is presented.

Keywords: Segmentation, K-means Clustering, Life Insurance.

1. Introduction and preliminaries

One of the most important existing insurance lines of business in the world is considered to be life insurance. This is such that in countries with developed and leading insurance industries, insurers are classified into two distinct categories namely; life and non-life insurers.

Although this field has a high penetration rate in developed countries, in Iran we have not yet reached the desirable situation. Comparing to the past, life insurance has been modified in such a way that there are currently products that can fit into any budget. Furthermore many times people do not even have to pass physical exams.

*Corresponding author
Email addresses: khamesian@irc.ac.ir (Farzan Khamesian), khanizadeh@irc.ac.ir (Farbod Khanizadeh), alireza.bahiraie@semnan.ac.ir (Alireza Bahiraie)

Received: January 2021 *Accepted:* February 2021

In fact proper marketing plays a crucial role to provide services meeting the customer needs. Marketing in one way may be translated to gaining new clients for insurance companies. However in long term the marketing strategies will be efficient only if the clients are happy with the services. Therefore it is essential to have a good understanding over the need and behaviors of customers.

This leads us to the concept of customer segmentation [25, 18] providing insight into the target market. By means of segmentation companies can group prospective client into different segments sharing common characteristics. Some examples of customer segmentation can be found in [22, 11, 1, 20].

Recently Data mining techniques have extensively exploited by different sectors and enterprises to develop more effective marketing strategies; see for example [21, 15, 9, 16, 5]. Regarding market segmentation, it is also known as clustering and the techniques used to develop the models are called clustering algorithms [12, 3, 24].

2. METHODOLOGY (K-Means Clustering)

Clustering is a general technique which most people experience in their life. We can come up with plenty of situations resembling clustering concept. For example consider the case where you try to sort your books based on their themes and subjects. Also restaurants in which group of friends or family members are sitting around a table or items arranged in a mall are instances of clustering. Some big compaies such as Amazon and Netflix set their recommendation systems based on clustering [2, 14].

In fact clustering is the process of dividing the datasets into groups, consisting of similar data-points. Therefore points in the same group are as similar as possible whereas points belonging to different clusters are significantly dissimilar to each other [14]. In other words in clustering, data are organized into groups such that there is high intra-cluster similarity and low inter-cluster similarity.

Clustering is an unsupervised approach in machine learning [26] and as mentioned above it is built upon the notion of similarity/dissimilarity. Working with the notion of dissimilarity, two data points are considered 'close' if their dissimilarity or distance is small. However when we have the concept of similarity, two individuals are close when their similarity is large enough [10]. Therefore we need to define similarity/dissimilarity metric for data points. For different types of data different metrics are introduced [7, 23, 6, 17]. Distance measure is another jargon used to define similarity metrics. In fact similarity reflects the level of relationship between two data items, while dissimilarity assesses the measurement of divergence between two data items. Appropriate distance function can increase the performance of clustering algorithms(Figure 1) [4, 19]:

A general term encompassing both similarity and dissimilarity is called proximity. A common approach toward proximity measures start with construction of the so-called data matrix in which columns and rows present number of features and observations respectively [13, 8]:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{21} & \cdots & \cdots & x_{21} \end{pmatrix}$$

In each row one can find information about a distinct observation (i.e. object). So $x_1 = (x_{11}, x_{12}, \ldots, x_{1p})$ stands for the first record in the data set and each component represents a particular variable. Therefore in general the entry x_{ij} in the matrix provides the value of the jth variable on observation i. This matrix is also known as 'two-mode' since the rows and columns indicate different things.

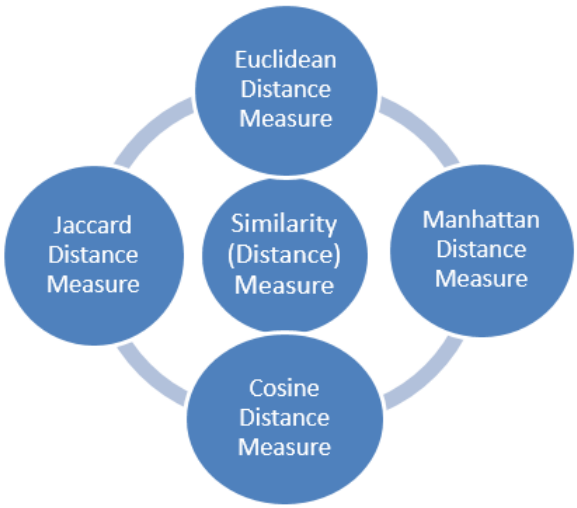

Figure 1: Examples of Similarity Measures

		Observation i		
	Outcome	1	0	Total
Observation j	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

Figure 2: Number of binary outcomes for two observations

Another common matrix that is used in clustering analysis is the n by n matrix written in the form of:

$$\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{pmatrix}$$

This is the proximity matrix with entries denoting proximity values for all pairs of n observations. In fact $d(i,j)$ calculates the dissimilarity/distance between two data points. In case it equals zero we face the lowest dissimilarity and consequentlt the highest similarity between points. Therefore the relationship between similarity and dissimilarity can be expressed as:

$$s_{ij} = 1 - d(i,j)$$

Note that different proximity measures are suitable for different types of data such categorical, numerical and mixed data. In case we are dealing with categorical variables of two states (i.e. binary variables: 0 or 1) the tables (Figure 2) could provide all possible combinations:

In the table shown in Figure 2, a is the number of features that take the value 1 for both observations i and j, b presents the number of those variables taking 0 and 1 for ith and jth observation respectively, c is the number of features equal 1 for observation i and 0 for observation j, and finally d is the number of columns taking the value 0 for both ith and jth observations. Clearly we have p

Measure	Formula
Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
Jaccard coefficient	$s_{ij} = a/(a+b+c)$
Rogers and Tanimoto	$s_{ij} = (a+d)/[a+2(b+c)+d]$
Sneath and Sokal	$s_{ij} = a/[a+2(b+c)]$
Gower and Legendre	$s_{ij} = (a+d) \Big/ \left[a+\frac{1}{2}(b+c)+d\right]$
Gower and Legendre	$s_{ij} = a \Big/ \left[a+\frac{1}{2}(b+c)\right]$

Figure 3: Examples of Proximity Measures for Categorical Variables.

features which can be written as $p = a + b + c + d$. In this case different similarity measures can be defined as provided in Figure 3.

When the variables are continuous the proximity between two observations are mainly described based on the distance between two points. In geometrical sense Euclidean distance is the most standard approach. Consider two n-dimensional vectors as:

$$x_1 = (x_{11}, x_{12}, \cdots, x_{1n})$$
$$x_2 = (x_{21}, x_{21}, \cdots, x_{2n})$$

Then the Euclidean distance is expressed as the root of square differences between the coordinates of each pair:

$$d_E = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \cdots + (x_{1n} - x_{2n})^2}$$

Manhattan distance is the other metric to obtain the distance between two points. In this case for a pair of two n-dimensional data points the distance becomes the sum of absolute difference between individual components:

$$d_M = |x_{11} - x_{21}| + |x_{12} - x_{22}| + \cdots + |x_{1n} - x_{2n}|$$

Note that distance functions follow the following properties:

$$d(a,b) \geq 0 d(a,b) = d(b,a) d(a,b) \leq d(a,c) + d(c,b)$$

The different graphical presentation of Manhattan and Euclidean can be see in the following plots. Euclidean distance moves along the shortest straight line between two data points while for Manhattan distance the so called block path is plotted as Figure 4.

The other distance measure is the so called cosine metric. This measure determines the angle between two vectors given by:

$$d_{\cos} = 1 - \frac{x_{11}x_{21} + x_{12}x_{22} + \cdots + x_{1n}x_{2n}}{\left(\sqrt{x_{11}^2 + x_{12}^2 + \cdots + x_{1n}^2}\right)\left(\sqrt{x_{21}^2 + x_{22}^2 + \cdots + x_{2n}^2}\right)}$$

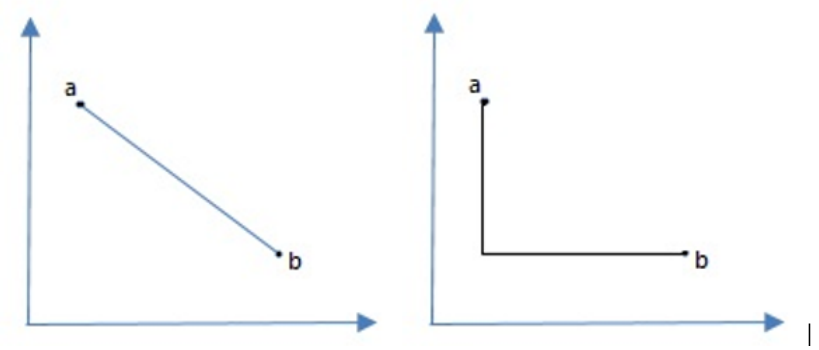

Figure 4: left: Euclidean right: Manhattan

Basically cosine measure determines the proximity between two vectors through cosine of the angle between given vectors in their dot product space. Note that the vector length does not affect the measure value. This makes it a proper choice for high-dimensional data sets. The above expressions present the case for two data points. The table 5 provides some proximity measures in general form in the case of continuous variables in p-dimensional space. All distance measures are written with weighting factor w_k assessing the weights of the p variables.We commonly set $w_k = 1$, however the complete description of obtaining weight can be found in [10].

3. EMPIRICAL RESULTS

3.0.1. Developing the Model

Data set for this study is collected from an Iranian insurance company containing around 5000 records of customers who have purchased life insurance. There are five variables namely; age, gender, TPL coverage limit, premium paid for car insurance other than TPL and premium for life insurance.

The aim is to group customers into clusters sharing similar characteristics regarding car and life insurance premium. In fact people in the same cluster demonstrate homogeneous behavior toward life insurance policy. However the main challenge of K-means clustering is to determine the optimal number of clusters (i.e. number of K). With the help of elbow method we would run clustering on the given data set for ten different values of $K(i.e. K = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$. Figure 6 provides the minimum efficient number of cluster as $K = 4$.

As can be seen in Figure 6 the most efficient performance of the model appears for $K = 4$. This means the data points can be divided into 4 different clusters in which individuals share similar patterns. Running K-means clustering for K equals 4 we have the Figure 7.

Figure 7 demonstrates the relationship between car insurance policies and the willingness to apply for life insurance. In fact, Figure 7 shows the link between costumer revenue and the demand for life insurance policies. The horizontal axis contains 3 items defining changes in budgets associated with car insurance. The first one is the maximum coverage for third party liability (TPL) car insurance (i.e. Max Cover). This can result in an increase in the cost of car insurance. The second one is the number of different car policies purchased by clients rather than TPL (e.g car body insurance). The last element that contributes greatly to the cost of car insurance is the price level of the cars.

Since the third item highly depends on the car prices, by the growth of car insurance expenses one can guess an increase in the level of customers income. This fact, along with the two main factors namely age and gender of policyholders determine the difference between groups and provide a significant interpretation over the clusters which is the association between car and life insurance customers.

Measure	Formula
Euclidean Distance	$d_{ij} = \left[\sum_{k=1}^{p} w_k^2 \left(x_{ik} - x_{jk}\right)^2\right]^{1/2}$
City-Block (Manhattan) Distance	$d_{ij} = \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk} \rvert$
Minkowski Distance	$d_{ij} = \left(\sum_{k=1}^{p} w_k^r \lvert x_{ik} - x_{jk} \rvert^r\right)^{1/r} \quad (r \geq 1)$
Canberra Distance	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk} \rvert / (\lvert x_{ik} \rvert + \lvert x_{jk} \rvert) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
Pearson Correlation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot}) \Big/ \left[\sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})^2 \sum_{k=1}^{p} w_k \left(x_{jk} - \bar{x}_{j\cdot}\right)^2\right]^{1/2}$ where $\bar{x}_{i\cdot} = \sum_{k=1}^{p} w_k x_{ik} \Big/ \sum_{k=1}^{p} w_k$
Angular Seperation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2\right)^{1/2}$

Figure 5: Examples of Proximity Measures for Continuous Variables.

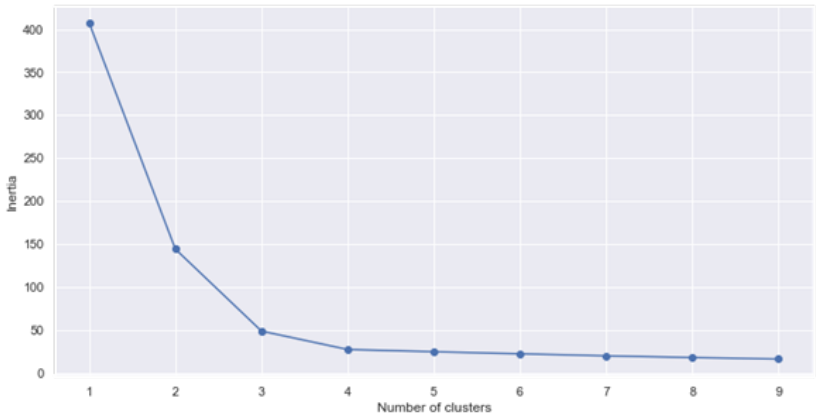

Figure 6: Optimizing the number of clusters.

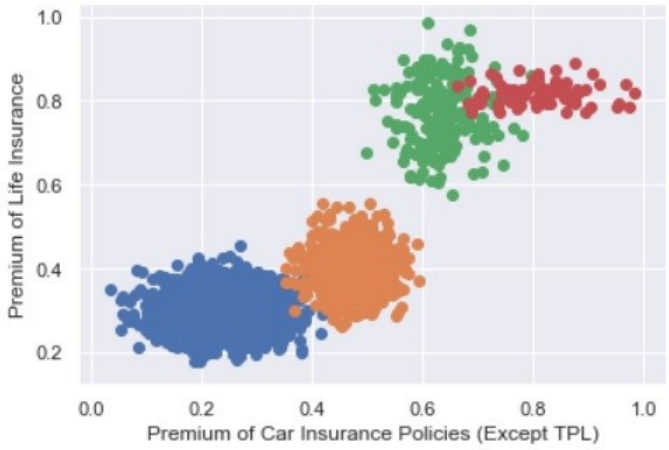

Figure 7: Clustering for $K = 4$.

			Percentage
Cluster 1	Gender	Female	31
		Male	69
	Age	Young	32
		Middle Age	41
		Old	27
	Max Cover	---	32
	Car Body	---	32

Figure 8: 1^{st} Cluster Statistics

Tables 8, 9, 10 and 11 present the descriptive distribution in four distinct clusters.

In the next session we see how Figure 7 and the above four tables can justify the pattern among customers holding both car and life policies.

3.1. Results

As reflected in Figure 7, the slope of the plot is positive which means there exist a positive correlation between the amount of money paid for car and life insurance premiums. This shows the relationship between the level of income and purchasing power of insurers in the demand for life and savings insurance (which is a type of supplementary insurance of social insurance).

First cluster in Figure 7 contains the lowest cost group both in the case of car insurance as well as life insurance policies. The main characteristics of this group is the high ratio of men to women and high average age among the group members. Putting these features together reflects the fact that the cluster includes people from the low income level of the society and at the same time the breadwinner of the family, which due to the low purchasing power, is naturally expected not to be interested enough in applying for life and savings insurance policies.This cluster contains the most populated group.

In cluster 2 the high ratio of women to men, points out the fact that although they spend more on car and life insurance, not being a breadwinner in the family (due to the two characteristics mentioned above) yields less steep slope comparing to the third cluster. In cluster 3, which possesses highest increasing slope in the level of demand for life insurance, there is a high average age and

			Percentage
Cluster 2	Gender	Female	56
		Male	44
	Age	Young	49
		Middle Age	38
		Old	13
	Max Cover	---	41
	Car Body	---	47

Figure 9: 2^{nd} Cluster Statistics

			Percentage
Cluster 3	Gender	Female	26
		Male	74
	Age	Young	28
		Middle Age	55
		Old	17
	Max Cover	---	63
	Car Body	---	78

Figure 10: 3^{rd} Cluster Statistics

			Percentage
Cluster 4	Gender	Female	31
		Male	69
	Age	Young	24
		Middle Age	54
		Old	22
	Max Cover	---	68
	Car body	---	75

Figure 11: 4^{th} Cluster Statistics

the significant male-to-female ratio. This cluster shows customers with medium to high income level of which majority would apply for the maximum coverage of TPL car insurance formin the most responsible group of individuals. In the fourth cluster, while we face a significant level of life insurance policies, it contains a much smaller population comparing to cluster 3. In fact although the cost of car body insurance in cluster 4 is much higher than the third one (i.e due to luxury cars), individuals are not willing to purchase maximum TPL car and life insurance. These results along with the probable wealthy situation of people, indicate they does not pay much attention to the benefits of life insurance policies as well as the maximum coverage of TPL car insurance.

4. DISCUSSION

As mentioned in the paper, penetration rate is one of the most strategic plans of all insurance companies. Therefore they have tried to modify and customize their products meeting the needs of different people. To achieve this goal it is vital to have some idea regarding the customers. In fact proper marketing can be a key point in attracting new customers or retaining the loyal ones. Within this regard market and customer segmentation provide a high level overview regarding behavior of clients. K-means clustering was carried out through the article and four different clusters were obtained containing customers with similar characteristics. Given the above, what can be considered as a significant result of the paper for the relationship between life and car insurance customers is the importance of the third cluster to focus on as one of the target groups on advertising and marketing process.

References

[1] Bayer, J., "Customer segmentation in the telecommunications industry." *Journal of Database marketing & customer strategy management*, **17**(3-4) (2010), pp.247-256.
[2] Bahiraei, A., Abbasi, B., Omidi, F., Hamzah, N. A., and Yaakub, A. H., "Continuous time portfolio optimization". *International Journal of Nonlinear Analysis and Applications*, **6** (2)(2014) pp. 103-112.
[3] Bahiraie, A., Azhar, A. K. M., and Ibrahim, N. A., "A new dynamic geometric approach for empirical analysis of financial ratios and bankruptcy." *Journal of Industrial & Management Optimization,* **7**(4) (2011), 947-965.
[4] Choi, S.S., Cha, S.H. and Tappert, C.C., "A survey of binary similarity and distance measures". *Journal of Systemics, Cybernetics and Informatics*, **8**(1) (2010), pp.43-48.
[5] Chiang, W.Y., "Applying data mining for online CRM marketing strategy." *British Food Journal*(2018).
[6] dos Santos, T.R. and Zárate, L.E., "Categorical data clustering: What similarity measure to recommend?." *Expert Systems with Applications,* **42**(3) (2015), pp.1247-1260.
[7] Everitt, B.S., Landau, S. and Leese, M., "Cluster Analysis," *4th edition. Edward Amold, New York I,* 993, (2001).
[8] Green, P.E. and Rao, V.R., "A note on proximity measures and cluster analysis", (1969).
[9] Han, J., Pei, J. and Kamber, M., "Data mining: concepts and techniques". *Elsevier*, (2011).
[10] Hastie, T., Tibshirani, R. and Friedman, J., " Unsupervised learning. In The elements of statistical learning." *Springer, New York, NY.* (2009) pp. 485-585.
[11] Hamka, F., Bouwman, H., De Reuver, M. and Kroesen, M., "Mobile customer segmentation based on smartphone measurement." *Telematics and Informatics*, **31**(2) (2014), pp.220-227.
[12] Holý, V., Sokol, O. and Černý, M., "Clustering retail products based on customer behaviour. Applied Soft Computing, **60** (2017), pp.752-762.
[13] Irani, J., Pise, N. and Phatak, M., "Clustering techniques and the similarity measures used in clustering: A survey". *International journal of computer applications*, **134**(7) (2016), pp.9-14.
[14] Linden, G., Smith, B. and York, J., "Amazon. com recommendations: Item-to-item collaborative filtering." *IEEE Internet computing,* **7**(1) (2003), pp.76-80.
[15] Linoff, G.S. and Berry, M.J., "Data mining techniques: for marketing, sales, and customer relationship management." *John Wiley & Sons*, (2011).
[16] Moro, S., Laureano, R. and Cortez, P., "Using data mining for bank direct marketing: An application of the crisp-dm methodology", (2011).

[17] Mori, U., Mendiburu, A. and Lozano, J.A., " Similarity measure selection for clustering time series databases." *IEEE Transactions on Knowledge and Data Engineering,* **28**(1) (2015), pp.181-195.
[18] Parsell, R.D., Wang, J. and Kapoor, C., Microsoft Corp, "Customer segmentation." *U.S. Patent Application*, 13/716,234, (2014).
[19] Patidar, A.K., Agrawal, J. and Mishra, N., "Analysis of different similarity measure functions and their impacts on shared nearest neighbor clustering approach." *International Journal of Computer Applications*, **40**(16) (2012), pp.1-5.
[20] Peker, S., Kocyigit, A. and Eren, P.E., "LRFMP model for customer segmentation in the grocery retail industry: a case study." *Marketing Intelligence & Plannin* (2017).
[21] Shaw, M.J., Subramaniam, C., Tan, G.W. and Welge, M.E., "Knowledge management and data mining for marketing." *Decision support systems*, **31**(1)(2001), pp.127-137.
[22] Teichert, T., Shehu, E. and von Wartburg, I., "Customer segmentation revisited: The case of the airline industry." *Transportation Research Part A: Policy and Practice*, **42**(1) (2008), pp.227-242.
[23] Torres, G.J., Basnet, R.B., Sung, A.H., Mukkamala, S. and Ribeiro, B.M., " A similarity measure for clustering and its applications." *Int J Electr Comput Syst Eng,* **3**(3) (2009), pp.164-170.
[24] Tsai, C.F., Hu, Y.H. and Lu, Y.H., "Customer segmentation issues and strategies for an automobile dealership with two clustering techniques." *Expert Systems,* **32**(1) (2015), pp.65-76.
[25] Weinstein, A.T., "Market segmentation: Using demographics, psychographics and other niche marketing techniques to predict customer behavior." *Probus Publishing Co.* (1994).
[26] Xu, R. and Wunsch, D., "Clustering." *John Wiley & Sons*, **10**(2008).