 ISSN: 2579-7298

Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method

Agus Perdana Windarto

STIKOM Tunas Bangsa, Jendral Sudirman Blok A No.1/2/3, Pematangsiantar, Indonesia
agus.perdana@amiktunasbangsa.ac.id

ARTICLE INFO

Article history:
Received: 2017-05-30
Revised: 2017-06-04
Accepted: 2017-08-22

Keywords:
Clustering
K-Means
Data Mining
Import Rice
Country of origin

ABSTRACT

Indonesia is a country where most of its people rely on the agricultural sector as a livelihood. Indonesia's rice production is so high that it cannot meet the needs of its population. Consequently Indonesia still has to import rice from other food producing countries. One of the leading causes is the enormous population. Statistics show that in the range of 230-237 million people, the staple food of all residents is rice, so it is clear that the need for rice becomes very large. This study discusses the application of data mining on rice import by leading country of origin using K-Means Clustering Method. Sources of data of this study were collected based on import declaration documents produced by the Directorate General of Customs and Excise. Also since 2015, import data also comes from PT. Pos Indonesia, records of other agencies at the border, and the results of cross-border maritime trade surveys. The data used in this study is the data of rice imports by country of origin from 2000-2015 consisting of 10 countries namely Vietnam, Thailand, China, India, Pakistan, United States, Taiwan, Singapore, Myanmar, and Others. Variable used (1) total import of rice (net) and (2) import purchase value (CIF). The data will be processed by clustering rice imports by leading country of origin in 3 clusters, ie high imported cluster, medium introduced cluster, and low import level cluster. The clustering method used in this research is the K-Means method. Centroid data for high import level clusters 7429180 and 2735452,25, Centroid data for medium import level clusters 1046359.5 and 337703.05 and Centroid data for low import level clusters 185559.425 and 53089.225. The result is an assessment based on rice import index with two high imported cluster countries namely Vietnam and Thailand, four medium-level groups of moderate import countries namely China, India, Pakistan and Lainya and four low imported cluster countries namely USA, Taiwan, Singapore, and Myanmar. The results of the research can be used to determine the amount of rice imported by the leading country of origin.

I. Introduction

Indonesia is a country where most of its people rely on the agricultural sector as a livelihood. Indonesia's rice production is so high that it cannot meet the needs of its population. Consequently, Indonesia still has to import rice from other food producing countries. One of the leading causes is the enormous population. Statistics show that in the range of 230-237 million people, the staple food of all residents is rice, so it is clear that the need for rice becomes very large[1].

Cluster analysis is a multivariant technique with the primary objective of grouping objects based on the characteristics they possess. Today, cluster analysis has been applied in many fields written in various studies and journals[2]. In the clustering method, the central concept is emphasized the iterative cluster center search, where the center of the cluster is determined by the minimum distance

of each data in the center of the group [3]. The data used in this study are based on import importation information documents produced by the Directorate General of Customs and Excise through the website https://www.bps.go.id. In this case, the researcher raised the topic of rice import by leading country of origin where the process of the method is clustering[4]. The results of the cluster can be used as input for the Indonesian state as a form of mapping of the leading country of origin. The mapping process can be clustered into 3 (three) clusters, i.e., the highest import production, medium import production and low import production[5].

Endang Sugiharti in his research conducted clustering of lecturers data related to the activities and their performance by the implementation and responsibility of the K-Means method[6]. Cluster in the research into the cluster Networking, Software Engineering, and E-Learning. The clustering method used in this research is the K-Means method[7].

K-means can also be defined as a Clustering method that is included in the partitioning approach. The K-Means algorithm is a centroid model[8]. Centroid mode is a model that uses centroid to create clusters. The centroid is the midpoint of a cluster. The centroid is a value. The centroid is used to calculate the distance of a data object against the centroid[9].

A data object is included in the cluster if it has the shortest distance to the cluster's centroid. K-Means algorithm can be interpreted as a simple learning algorithm to solve a grouping problem that aims to minimize double faults [10]. The purpose of this research is to apply K-Means in clustering rice imports by leading country of origin[11].

II. Method

In this study, the technique used is the method of data mining as follows. (A) Data collection stage, (b) Data processing stage, (c) Clustering stage and (d) Stage Analysis. Scenes in the method are further described as follows: In the application of data mining of imported rice production by the first country of origin, relevant data is required. Sources of research data obtained from the data collected by documents of importation of imports produced by the Directorate General of Customs and Excise through the site https://www.bps.go.id. The data used in this study is the data of rice imports by country of origin from 2000-2015 consisting of 10 countries namely Vietnam, Thailand, China, India, Pakistan, United States, Taiwan, Singapore, Myanmar, and Others. Variable used (1) the amount of rice import (net) and (2) the value of import purchases (CIF). The data will be processed by clustering rice imports based on central destination countries in 3 clusters, i.e., high imported cluster, medium introduced cluster, and low imported cluster level. The data that has been processed will be processed first to be clustered. In the previous stage, the data of each country of origin of rice imports will be summed in every aspect so that at this stage has been obtained the calculation of the value to be processed at the clustering stage. Clustering is an unattended classification and is a process of partitioning a set of data objects from one set into multiple classes. This can be done by applying various equations and steps about the distance of the algorithm, i.e., with Euclidean Distance. Cluster analysis is a method used to divide the data set into groups based on predetermined similarities. In determining the cluster based on the data already available, it takes a flowchart to facilitate in determining the flow of calculation as a groove to find the results of the application of the cluster to the data to be processed. Here is a flowchart in determining the cluster with K-Means. At this stage, the data analysis of rice imports by country of origin with the primary application of the tool. Rapid Minner. Rapid Miner is a machine learning environment for mining, text mining and predictive analytics. The data obtained is processed by using the weight calculation of each index. In the preceding stages, it has been determined to be clustered into 3 clusters of high imported clusters; medium introduced clusters and low import level clusters. At this stage will be analyzed the results.

III. Result and Discussion

In clustering, the data obtained will be calculated in advance based on the number of rice imports in 2000-2015 based on the leading country of origin. The sum result based on two assessment criteria is net weight and CIF value as shown in table 1.

Table 1. Rice Import Data by Main Country of Origin, 2000-2015

Country of Origin	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
	Netto : Ton															
Vietnam	369 546,8	142 511,8	561 728,9	506 012,8	58 810,1	44 772,5	272 832,7	1 022 834,6	125 070,5	20 970,5	467 369,6	1 778 480,6	1 084 782,8	171 286,6	306 418,1	509 374,2
Thailand	361 734,8	189 655,9	418 697,6	492 114,2	129 421,5	126 408,9	157 983,3	363 640,1	157 007,3	221 372,6	209 127,8	938 695,7	315 352,7	94 633,9	366 203,5	126 745,7
Tiongkok[1]	476 776,7	24 728,5	126 768,4	54 440,4	110,5	1,3	100,0	901,4	3 341,7	5 167,6	3 637,4	4 674,8	3 099,3	639,8	1 416,7	479,9
India	0,1	2 047,1	405 032,2	108 797,4	923,3	327,0	720,6	3 571,8	289,5	473,1	601,3	4 064,6	259 022,6	107 538,0	90 653,8	34 167,5
Pakistan	20 139,1	26 110,0	32 281,4	49 071,0	0,0	0,0	904,3	4 603,6	751,3	501,5	4 992,1	14 342,3	133 078,0	75 813,0	61 715,0	180 099,5
Amerika Serikat	49 405,2	177 889,1	13 392,9	107 607,6	16 766,7	2 184,2	801,0	821,7	1 411,2	1 323,4	1 644,1	2 074,1	2 445,5	2 790,4	1 078,6	0,0
Taiwan	0,0	0,0	3 541,5	9 600,5	10 600,0	0,0	2 500,0	625,0	0,0	0,0	0,0	5 000,0	0,0	1 240,0	840,0	0,0
Singapura	2 496,9	7 268,7	21 809,7	4 314,8	6 614,8	6 632,1	1 568,4	468,3	898,2	250,0	10,8	1 506,5	22,5	0,5	0,0	0,0
Myanmar	198,4	25 441,3	111 687,3	41 398,9	2 500,0	0,0	0,0	0,0	0,0	0,0	0,0	1 140,0	11 819,6	18 450,0	15 616,0	8 775,0
others	75 367,9	49 080,6	110 439,9	55 148,2	11 119,8	9 290,6	698,3	9 381,2	919,8	414,5	198,4	497,8	749,3	272,5	222,0	1 959,2
amount	1 355 665,9	644 732,8	1 805 379,9	1 428 505,7	236 866,7	189 616,6	438 108,5	1 406 847,6	289 689,4	250 473,1	687 581,5	2 750 476,2	1 810 372,3	472 664,7	844 163,7	861 601,0
	Nilai CIF: 000 US$															
Vietnam	77 075,6	23 440,6	112 965,1	96 273,1	13 544,1	12 387,4	83 773,4	335 558,8	47 392,1	7 936,9	232 915,7	946 490,1	564 925,7	97 303,3	143 536,0	202 563,1
Thailand	83 408,0	33 756,6	82 566,1	110 447,0	33 023,2	32 489,0	45 399,3	122 425,9	64 721,0	81 959,8	109 133,7	533 001,9	186 171,4	61 787,5	175 387,4	66 772,4
Tiongkok[1]	107 428,4	4 645,6	25 623,5	11 131,6	42,5	4,0	145,0	1 500,6	6 642,4	13 697,0	12 728,5	15 467,1	11 205,6	1 526,5	4 101,5	1 631,0
India	0,1	303,7	61 148,0	17 282,7	423,1	317,7	644,2	1 606,2	337,6	791,5	1 767,5	6 307,9	122 189,0	44 989,1	34 299,5	13 671,7
Pakistan	5 231,1	5 045,8	5 464,8	8 198,1	0,0	0,0	188,2	1 247,0	259,2	160,7	1 765,8	6 053,4	52 483,4	29 996,9	23 909,3	62 949,2
Amerika Serikat	13 828,9	52 888,6	3 579,3	22 830,5	5 459,6	778,7	646,9	791,1	1 796,3	2 005,7	1 745,5	2 489,6	2 718,6	2 983,6	1 294,3	0,0
Taiwan	0,0	0,0	521,6	3 358,1	3 786,5	0,0	935,6	223,6	0,0	0,0	0,0	1 050,0	0,0	465,6	252,0	0,0
Singapura	484,3	1 405,4	3 672,8	821,4	1 698,4	1 880,3	516,2	170,9	465,5	100,0	27,6	981,9	32,2	1,4	0,0	0,0
Myanmar	13,8	3 822,2	16 969,3	6 502,0	427,1	0,0	0,0	0,0	0,0	0,0	0,0	419,2	4 754,5	6 498,9	5 082,8	2 732,3
others	31 659,8	9 604,2	30 016,5	14 578,3	3 348,3	3 641,9	371,7	4 195,3	2 528,8	1 501,6	700,8	902,3	1 142,8	449,2	315,7	1 282,5
amount	319 130,0	134 912,8	342 527,1	291 422,9	61 752,8	51 499,0	132 620,5	467 719,4	124 142,8	108 153,3	360 785,0	1 513 163,5	945 623,2	246 002,1	388 178,5	351 602,2

The data is then accumulated based on two criteria, i.e., net weight and (2) CIF value as shown in table 2 below:

Table 2. Rice Import Accumulation Data

No	Country of Origin	Berat Netto(Ton)	Nilai CIF (US$)
1	Vietnam	8973929,9	3336257,3
2	Thailand	5884429,9	2134647,2
3	Tiongkok1	1582007,8	463422,3
4	India	1019299,9	305743,5
5	Pakistan	732306,3	234729
6	Amerika Serikat	455790,5	154904,5
7	Taiwan	70968,5	20237,8
8	Singapura	3156,8	1779,5
9	Myanmar	212321,9	35435,1
10	Lainnya	851824	346917,4

Once accumulated it will get the value of all rice imports by the destination country. Then the data will go into the clustering stage by applying the K-Means algorithm using rapid manner to cluster the data into three clusters. Accumulated data will be entered into the Rapid Minner tool. So it can be clustered into two. Once introduced into Rapid Minner. The results of data accumulation can be seen in table 2. In the application of the K-means algorithm the value of midpoint or centroid is obtained from the data collected with the provision that the desired clusterization is 3, the cluster determination is divided into three parts namely the high import level cluster (C1), the medium import cluster (C2) and the cluster level Low import (C3). Then the value of the midpoint or centroid also there is 3 points. The determination of the cluster point is carried out by taking the largest (maximum) amount for the high imported level cluster (C1), the average cost for the medium imported cluster (C2) and the smallest (minimum) amount for the low import level cluster (C3). The point value can be seen in Table 3 below:

Table 3. Centroid Initial Data

Atribut	**High import clusters**	
Value	Netto	CIF
	8973929,9	3336257,3
	Middle import clusters	
	Netto	CIF
	1582007,8	463422,3
	Low import clusters	
	Netto	CIF
	3156,8	1779,5

By using the centroid can be clustered data that has been obtained into 3 clusters. Cluster process by taking the closest distance from any data that is processed. From the import data of rice of the leading country of origin, the clustering was obtained in the iteration of 1 to 3 clusters. High import level clusters (C1), i.e. Vietnam and Thailand, medium imported clusters (C2), i.e. China, India and Others and low imported cluster C3, i.e. Pakistan, USA, Taiwan, Singapore, and Myanmar. The process of finding the shortest distance, the data grouping in iteration 1 and Clustering data can be described in the following tables and figures:

Table 4. Calculation of cluster central gaps

No	**Country**	**Netto (Ton)**	**CIF (US$)**	**C1**	**C2**	**C3**	**Shortest Distance**
1	Vietnam	8973929,9	3336257,3	0	7930554	9570450	0
2	Thailand	5884429,9	2134647,2	3314948	4615607	6256077	3314948
3	Tiongkok[1]	1582007,8	463422,3	7930554	0	1644957	0
4	India	1019299,9	305743,5	8512353	584382	1060632	584382
5	Pakistan	732306,3	234729	8805898	879939	765457	765457
6	Amerika Serikat	455790,5	154904,5	9092838	1167711	477833	477833
7	Taiwan	70968,5	20237,8	9500458	1574691	70279	70279
8	Singapura	3156,8	1779,5	9570450	1644957	0	0
9	Myanmar	212321,9	35435,1	9362756	1434996	211855	211855
10	Lainnya	851824	346917,4	8654753	739420	916164	739420

Table 5. Grouping of iterative data 1

No	**Country**	**C1**	**C2**	**C3**
1	Vietnam	1		
2	Thailand	1		
3	Tiongkok[1]		1	
4	India		1	
5	Pakistan			1
6	Amerika Serikat			1
7	Taiwan			1
8	Singapura			1
9	Myanmar			1
10	Lainnya		1	

In table 5, the K-Means process will continue to iterate until the data grouping equals the previous iteration data grouping. In other words, the process will continue iterating until the data in the last iteration is the same as the previous iteration. The iterative grouping graph 1 can be seen in the following figure:

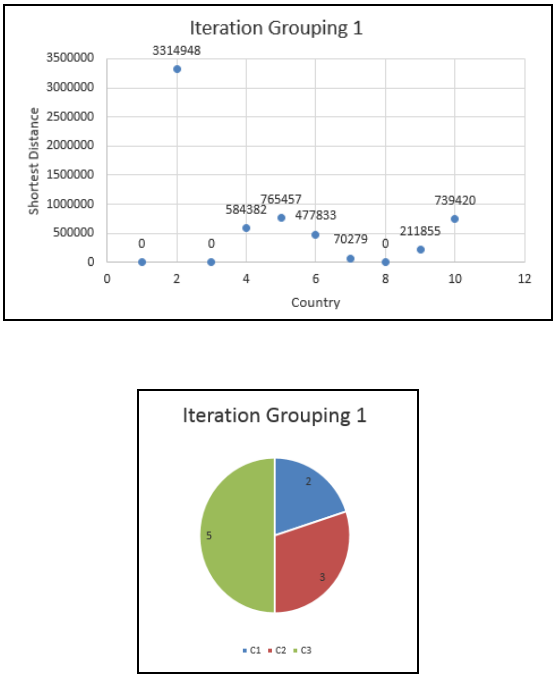


Fig. 1.Clustering of Iteration Data 1

In iteration 1, it is obtained clusters of rice import data of the main country of origin, which can be seen in Figure 2. In iteration 2, the process of midpoint or centroid value for iteration two can be seen in Table 6 below:

Table 6. Centroid Data iteration 2

Atribut	**High import clusters**	
Value	Netto	CIF
	7429179,9	2735452,25
	Middle import clusters	
	Netto	CIF
	1151043,9	372027,733
	Low import clusters	
	Netto	CIF
	294908,8	89417,18

After getting the value of the midpoint or centroid, the same process is done by finding the closest distance. The process of finding the shortest distance, the data grouping in iteration 2 and Clustering data can be described in the following tables and figures:

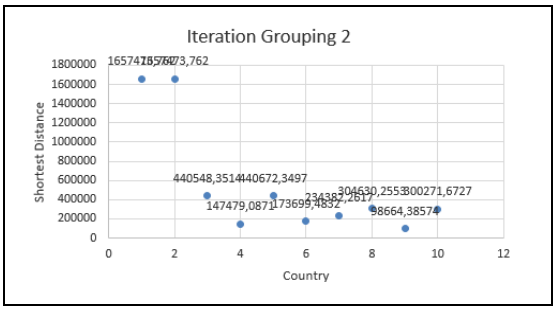
Table 7. Calculation of the center distance of the iteration cluster 2

No	**Country**	**Netto (Ton)**	**CIF (US$)**	**C1**	**C2**	**C3**	**Shortest Distance**
1	Vietnam	8973930	3336257	1657474	8365656	9266465	1657474
2	Thailand	5884430	2134647	1657474	5050918	5951950	1657474
3	Tiongkok[1]	1582008	463422,3	6273081	440548,4	1340337	440548,4
4	India	1019300	305743,5	6854929	147479,1	756002,3	147479,1
5	Pakistan	732306,3	234729	7148548	440672,3	460903,6	440672,3
6	Amerika Serikat	455790,5	154904,5	7435549	728367,9	173699,5	173699,5
7	Taiwan	70968,5	20237,8	7843192	1135922	234382,3	234382,3
8	Singapura	3156,8	1779,5	7913203	1206121	304630,3	304630,3
9	Myanmar	212321,9	35435,1	7705396	997243	98664,39	98664,39
10	Lainnya	851824	346917,4	6997622	300271,7	613564,1	300271,7

Table 8. Grouping of iterative data 2

No	Country	C1	C2	C3
1	Vietnam	1		
2	Thailand	1		
3	Tiongkok[1]		1	
4	India		1	
5	Pakistan		1	
6	Amerika Serikat			1
7	Taiwan			1
8	Singapura			1
9	Myanmar			1
10	Lainnya		1	

From table 8, the result of the 2nd iteration grouping is obtained with unequal results in the iteration 1. The process will take place in the next iteration. The iterative grouping graph 1 can be seen in the following figure:



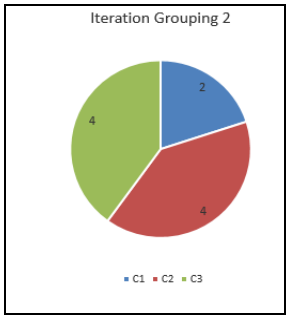


Fig. 2.Clustering of Iteration Data 2

Based on figure 2, from the import data of rice of the main country of origin, the clustering of iterations 2 for 3 clusters is obtained. High import level clusters (C1) ie Vietnam and Thailand, medium-level import clusters (C2) ie China, India Pakistan and Others and imported low-level clusters (C3) ie the United States, Taiwan, Singapore and Myanmar. The result of iteration 2 shows different results with the result of iteration 1. Then the process will proceed to iteration 3 with the same calculation that determines the new centroid based on iteration 2 and find the shortest distance from the centroid value on iteration 3

Table 9. Centroid Data iteration 3

Atribut	High import clusters	
	Netto	CIF
Nilai	7429179,9	2735452,25
	Middle import clusters	
	Netto	CIF
	1046359,5	337703,05
	Low import clusters	
	Netto	CIF
	185559,425	53089,225

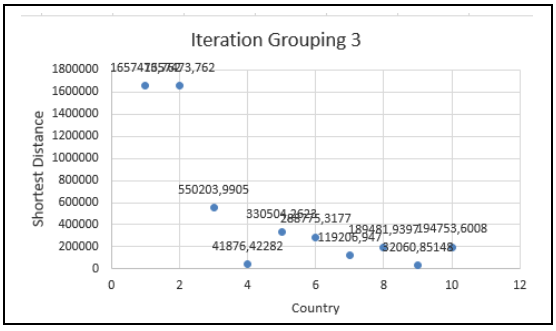
After getting the value of the midpoint or centroid, the same process is done by finding the closest distance. The process of finding the shortest distance, data grouping on iteration 3 and Clustering data can be described in the following tables and figures:

Table 10. Calculation of the center distance of the iteration cluster 3

No	Country	Netto (Ton)	CIF (US$)	C1	C2	C3	Shortest Distance
1	Vietnam	8973930	3336257	1657474	8475712	9381612	1657474
2	Thailand	5884430	2134647	1657474	5161001	6067125	1657474
3	Tiongkok[1]	1582008	463422,3	6273081	550204	1455487	550204
4	India	1019300	305743,5	6854929	41876,42	871181,6	41876,42
5	Pakistan	732306,3	234729	7148548	330504,3	576129,5	330504,3
6	Amerika Serikat	455790,5	154904,5	7435549	618212,8	288775,3	288775,3
7	Taiwan	70968,5	20237,8	7843192	1025754	119206,9	119206,9
8	Singapura	3156,8	1779,5	7913203	1095955	189481,9	189481,9
9	Myanmar	212321,9	35435,1	7705396	887121,5	32060,85	32060,85
10	Lainnya	851824	346917,4	6997622	194753,6	728178,2	194753,6

Table 11. Grouping of iterative data 3

No	Negara	C1	C2	C3
1	Vietnam	1		
2	Thailand	1		
3	Tiongkok[1]		1	
4	India		1	
5	Pakistan		1	
6	Amerika Serikat			1
7	Taiwan			1
8	Singapura			1
9	Myanmar			1
10	Lainnya		1	



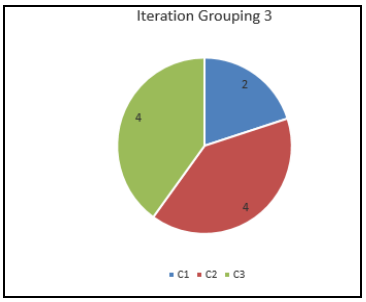


Fig. 3.Clustering of Iteration Data 3

In iteration 3, the grouping of data conducted on 3 clusters with iteration 2 obtained the same result. Of the 10 main rice import data of the leading country of origin can be known, 2 Cluster of high import level (C1) ie Vietnam and Thailand, 4 clusters of moderate import level (C2) ie China, India Pakistan and Others and 4 clusters of low import level (C3) United, Taiwan, Singapore and Myanmar.

IV. Conclusion

To assess the import of rice the leading country of origin can be Is done by applying a K-Means clustering method. The data is processed to derive the value of the imported rice production of the leading country of origin. The data were processed using Rapid miner to determine centroid values in 3 clusters, i.e., high introduced cluster level (C1), medium imported clusters (C2) and low imported level clusters (C3). Centroid data for imported high-level clusters 7429179,9 and 2735452,25, Centroid data for medium import level clusters 1046359.5 and 337703.05 and Centroid data for low import level clusters 185559.425 and 53089.225. Hence, it was obtained the assessment based on rice import index with 2 clusters of high import level (C1), i.e., Vietnam and Thailand, four medium imported clusters (C2), ie China, India Pakistan and Others and 4 clusters of low import level (C3), ie USA, Taiwan, Singapore, and Myanmar. The results of the research can be used to determine the amount of rice imported by the leading country of origin. In the processing of data for clustering can provide weighting criteria for data produced more accurate. Also, it also needs to increase the accuracy of clustering on the data. Added rules can be made to get more accurate results in the clustering process.

Acknowledgment

Thanks to the information system study program, STIKOM Tunas Bangsa Pematangsiantar for support in conducting this research.

References

[1] D. Mining, “Springer Series in Statistics The Elements of,” *Math. Intell.*, 2009.
[2] X. Wu, X. Zhu, G. Q. Wu, and W. Ding, “Data mining with big data,” *IEEE Trans. Knowl. Data Eng.*, 2014.
[3] C. C. Aggarwal, *Data Mining: The Textbook*. 2015.
[4] A. Vinnikov and S. Shalev-Shwartz, “K-means recovers ICA filters when independent components are sparse,” *Proc. 31st Int. Conf. Mach. Learn.*, 2014.
[5] C. Romero and S. Ventura, “Data mining in education,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2013.
[6] K. He, F. Wen, and J. Sun, “K-means hashing: An affinity-preserving quantization method for learning binary compact codes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013.
[7] G. Tzortzis and A. Likas, “The MinMax k-Means clustering algorithm,” in *Pattern Recognition*, 2014.
[8] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, “An improved overlapping k-means clustering method for medical applications,” *Expert Syst. Appl.*, 2017.
[9] S. Na, L. Xumin, and G. Yong, “Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm,” in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010.
[10] D. Boley *et al.*, “Spectral Rotation versus K-Means in Spectral Clustering,” *Twenty-Seventh AAAI Conf. Artif. Intell.*, 2013.
[11] X. Cai, F. Nie, and H. Huang, “Multi-view K-means clustering on big data,” in *IJCAI International Joint Conference on Artificial Intelligence*, 2013.