International Journal of Engineering & Technology, 7 (4.44) (2018) 43-46

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Clustering Analysis of Premier Research Fields

Terttiaavini[1], Fakhry Zamzam*[2], Mustafa Ramadhan[1], Azrai'ie K. Rosni[2], Tedy Setiawan Saputra[2], Agustina Heryati[1] and Dhamayanti[1]

[1] Computing Faculty Indo Global Mandiri University, Palembang -Indonesia
[2] Economic Faculty, Indo Global Mandiri University, Palembang - Indonesia
**Corresponding author Email: Fakhry@uigm.ac.id*

Abstract

The clusterization is one of methods which utilized to grouping a dataset which has a specific characteristics value. The processed data can be numerical or non-numerical data. Non-numeric data must be transformed first into numerical data. The case study in this study was to group research from six fields of science. The research data is non-numerical data is converted into the research contributions percentage in the science field. Utilized the c-means algorithm, the data was successfully grouped into three excellent research fields. The aim of the clustering is to know how many researchers in one cluster. Dataset is processed by utilizing the c-means algorithm to generated 3 clusters, they are an expeditious technology, entrepreneur and economic creative development, social engineering and strategic area infrastructure development. The data clustering result is presented in the graphic form by utilized the studio Rapidminer application.

Keywords: *Clustering, Premier research fields, C-Means algorithm, Euclidean distance*

1. Introduction

Currently, the data mining has been widely applied to solving any various problems [1] together with the growth of the information technology, every organization already owned a large data storage to support all of the activities. The data will be increasing and will become a trash if not reuse. Data can be processed to be a pattern and generated a new knowledge/ information. That data processing system is called data mining. Indo Global Mandiri University has a lecturer research data that has never been used. The number of lecturer at Indo Global Mandiri University is 110 lecturers. Every lecturer must carry out the research activities at least one time per semester. The research team can be consist of the 2-4 researcher with a different skill (joint research). The utility for data on how many lecturer research which supported a certain premier research, is difficult to answer. The development of the research field for the next five years is to mapping the lecturer research into three university premier research fields. The three university premier research fields are information technology, entrepreneurship and create economic development, social engineering, strategic area, and infrastructure development. To accelerate the process of providing data required an efficient way to classify the research data. One technique could be applied is the clustering the research data. The clustering method is a technique or method to classify data from the large one into the cluster which has similarities to the certain characteristics. The algorithm clustering method which applied is the c-means algorithm. It is able to group the categorical data and generated clusters that more stable with a short computation time. This research aims to classify the research data into the three university premier research fields with utilizing the clustering methods. This grouping is useful for mapping the lecturer research roadmap at Indo Global Mandiri University

2. Related work

The c-means algorithm is the algorithm clustering which most easy to apply on the small dataset [2].The c-means Algorithm Clustering can be applied with various techniques in the database [3] which are sourced from the multiple data sources [4]. Utilizing the c-means clustering model by determining a random initial centroid [5], determine the distance between objects and normalize the data to improve the process of c-means clustering [6]. The c-means algorithm has been applied in some research, such as: (1) Hygiene : Clustering of the Parkinson'sdisease [7] [8] ,Obese management [9], Health care knowledge discovery [10]; (2) Clustering image : Satellite Image [11], Segmentation of white blood cells [12], Brain image segmentation [13], Content based image retrieval (CBIR) [14], Banana Image Segmentation [15], Hand gesture segmentation [16], Segmentation of fruits based on color features [17]; (3) Network science: Network partition [18], Wireless sensor networks [19]; (4) Academic science : Student careers [20], Predicting students Performance [21] ; (5) Customer satisfaction : Evaluate the cluster customers [22], Customer satisfaction in fast-food restaurant [23]; (6) Multimedia applications [24]; (7) Chemical oxygen demand [25]; (7) Approach to characterize road accident locations [26]; (8) Watershed classification [27]; (9) Wind speed [28]; (10) Tax based on cluster [29]; (11) Plagiarism detection System [30]; (12) Dictionary learning [31]; (13) Crime analysis [32]; (14) Connection oriented telecommunication data [33]; (15) Analyze Software Architecture [34]; (16) Prediction of atomic web services reliability [35] etc. It has shown that the c-mean algorithm already implemented cases to solve the human problems

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

3. The Research Methodology

In this research, the first phase is a data pre-processing. The research data set consists of the research titles and five fields of science. The science field is a represented the expertise of the researchers. The research could be supported by several fields of science. The Data which represented the science field is transformed into the researcher contribution percentage on the research activities. The contribution percentage value is determined by the researcher team. Data is examined by utilizing the c-means algorithm with several iterations until it reaches convergent. The examination results show the clusters for each data. The data pattern which formed could be utilized as a cluster determination model for new data.

4. Data Pre-processing

The Preprocessing is an activity/process to change the original data into quality data that can be used for the next process. Before being processed utilizing the c-means algorithm, the data should no longer contain a missing value, value distortion or misrecording. The data sets must be clean up, integrating, reducing, adding data or transformation [4]. The un numeric data must be transformed into numerical, binary, nominal or scale. This study utilized a 20 data set. The researcher contribution percentage value on the study was determined by the research team. The stages of solving problems with the clustering method are described as follows:

4.1. Collecting data

The sample is a research data set from the Indo Global Mandiri University research institute repository which taken in 2018. The number of data which represents the sample is 20 research data. The science field is grouped into six science field are computer science, economics, engineering, government science, graphic design science, and linguistics. The science field is converted into the numerical which is the value of the researchers' contribution to the research activities. The total value of the contributions for a research is 100. One research could consist 1-4 the science field (collaboration research).

Table 1 : Research data conversions

Y	X1	X2	X3	X4	X5	X6
1	50	40	0	0	10	0
2	60	30	10	0	0	0
3	60	0	0	0	10	30
4	0	10	90	0	0	0
5	100	0	0	0	0	0
6	0	100	0	0	0	0
7	0	0	100	0	0	0
8	0	10	0	75	0	15
9	0	50	0	0	10	40
10	50	50	0	0	0	0
11	10	90	0	0	0	0
12	0	30	0	0	0	70
13	50	20	0	20	0	10
14	0	70	0	10	10	10
15	50	20	10	20	0	0
16	0	0	0	90	10	0
17	30	0	0	0	50	20
18	0	0	0	0	0	100
19	60	10	30	0	0	0
20	0	0	0	0	100	0

Variable Y represents the research, variable Xn represents the science field, where X1 is Computer Science, X2 is Economics, X3 is an Engineering science, X4 is a Graphic design science, X5 is a Government science of and X6 is a Language. I do not have any certain science element, will be given a zero.

4.2. Running C-Means algorithm

Clustering is the data mining method which does not require any unsupervised data. The Clustering is divided data sets into several parts (groups) that have similar characteristics. The C-means algorithm is a clustering algorithm that processes data repeatedly until it reaches a converging. Every repetition will calculate the center value of the cluster (centroid) and the distance of each data with the centroid. Then each data is classified based on its proximity to the centroid. The steps to solve the problem by utilizing the C-means algorithm are as follows:

1. *Phase 1: Determine the number of clusters*
   In the iteration-1, the researchers determined 3 initial centroids randomly from the research data set. The selected Centroids are at positions Y5 = 5, C0 = (100, 0,0,0,0,0), Y = 6, C1 = (0,100,0,0,0,0) Y = 20, C2 = (0,0,0,0,100,0). Given a grey sign-in table 1.

2. *Phase 2: Calculate the distance of each object to the centroid*
   Every object has calculated the distance to the centroid between utilizing the Euclidean distance (d).

3. *Phase-3: Determined the cluster for each object*
   The clusters of each object are determined based on the closest distance. The closest distance is the minimum value of the cluster for each object. Example: The closest distance object Y1 = 64.81 then the cluster for object Y1 = C0.

Table 2 : Determination of clusters for each object based on closest distance (cd)

Y	C1	C2	C3	cd	Cluster
1	64.81	78.74	110.45	64.81	C0
2	50.99	92.74	120.83	50.99	C0
3	50.99	120.83	112.25	50.99	C0
4	134.91	127.28	134.91	127.28	C1
5	0.00	141.42	141.42	0.00	C0
6	141.42	0.00	141.42	0.00	C1
7	141.42	141.41	141.42	141.42	C0
8	126.29	118.11	126.29	118.11	C1
9	119.16	64.81	110.45	64.81	C1
10	70.71	70.71	122.47	70.71	C0
11	127.28	14.14	134.91	14.14	C1
12	125.70	98.99	125.70	98.99	C1
13	58.31	96.95	115.76	58.31	C0
14	123.29	34.64	114.89	34.64	C1
15	58.31	96.95	115.76	58.31	C0
16	134.91	134.91	127.28	127.28	C2
17	88.32	117.47	61.64	61.64	C2
18	141.42	141.42	141.42	141.42	C0
19	50.99	112.25	120.83	50.99	C0
20	141.42	141.42	0.00	0.00	C2

4. *Phase-4: Calculated a new centroid*
   The new Centroid is determined based on object grouping for each cluster in one table. Cluster C0 consists of 10 objects, cluster C1 consists of 7 objects and C3 cluster consists of 3 objects. The centroid value of C0 is calculated based on the average coordinates of the C0 cluster. Likewise, calculations for centroid C1and C2. The results of the new centroid calculation are C0 = (68.80, 109.34, 124.26) C1=128.29, 65.42, 126.94) and C2 = (121.55, 131.27, 62.97). The next process is the same as phase 2 and phase 3.

5. *Phase-5: Compare new clusters with old clusters*
   The new cluster is compared to the cluster in the previous table. If there is a difference, it means that it has not reached the convergence. phase 2 and 3 are repeated until converging. The irritation process is done again by calculating the value of the new centroid.

5. Experimental result

The experiment result shows that to achieve a convergence in the sixth interaction with the C0=9 item for the entrepreneur and economic creative development group, C1=2 item for social engineering and infrastructure strategic area develoment, and C2= 9 for expeditious technology. Thus grouping base on the dataset which explained on table 3.

Table 3 : Clusterisation base on dataset

Y	X1	X2	X3	X4	X5	X6	Result
1	50	40	0	0	10	0	C2
2	60	30	10	0	0	0	C2
3	60	0	0	0	10	30	C2
4	0	10	90	0	0	0	C1
5	100	0	0	0	0	0	C2
6	0	100	0	0	0	0	C0
7	0	0	100	0	0	0	C1
8	0	10	0	75	0	15	C0
9	0	50	0	0	10	40	C0
10	50	50	0	0	0	0	C2
11	10	90	0	0	0	0	C0
12	0	30	0	0	0	70	C0
13	50	20	0	20	0	10	C2
14	0	70	0	10	10	10	C0
15	50	20	10	20	0	0	C2
16	0	0	0	90	10	0	C0
17	30	0	0	0	50	20	C2
18	0	0	0	0	0	100	C0
19	60	10	30	0	0	0	C2
20	0	0	0	0	100	0	C0

The data examine result utilized the Studio Rapidminer application, generated a scatter and pie on the charts, as below:

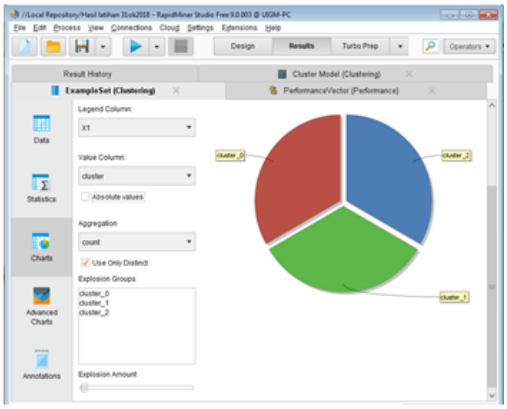


Fig 1 : Pie on the charts Display

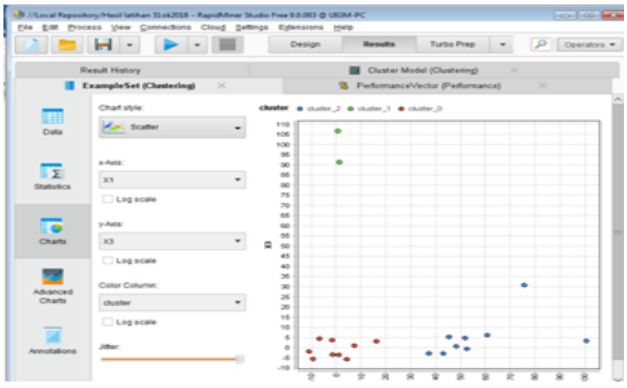


Fig 2 : Scatter on the charts Display

6. Conclusion

The c-means algorithm applicable on non-numerical datasets. The non-numerical data in this study are the lecturer research data at Indo Global Mandiri University. The data is converted into numerical data in the percentage contribution of the science field. The Grouping based on the six science fields of into three premieres research fields. The data reaches a convergent in the 6th iteration. This grouping is useful for determining the percentage of each research field for the development of the research field.

Acknowledgement

The research is funded by Directorate general strengthening research and development of the ministry of research, technology, and higher education through the decree number DIPA-042.06.1.401516/2018, with the higher education implemented research (PTUPT) scheme system

References

[1] K. Vadim, Overview of different approaches to solving problems of data mining, in: Procedia Comput. Sci., Elsevier B.V., 2018: pp. 234–239. http://doi:10.1016/j.procs.2018.01.036.

[2] K.A.A. Nazeer, M.P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, in: Proc. World Congr. Eng., 2009: pp. 1–5.

[3] E.M. Jane, E.G. Dharma, P. Raj, "SBKMMA : Sorting Based K Means and Median Based Clustering Algorithm Using Multi Machine Technique for Big Data", *Int. J. Comput.* Vol.28, (2018) pp:1–7.

[4] R. Wang, W. Ji, M. Liu, X. Wang, J. Weng, S. Deng, "Review on mining data from multiple data sources", *Pattern Recognit. Lett.* (2018) pp:1–9. http://doi:10.1016/j.patrec.2018.01.013.

[5] A.C. Fabregas, B.D. Gerardo, "Enhanced Initial Centroids for K-means Algorithm", *I.J. Inf. Technol. Comput. Sci.* Vol.1, (2017) pp:26–33. http://doi:10.5815/ijitcs.2017.01.04.

[6] N. Aggarwal, K. Aggarwal, Kirti, gupta, "Comparative Analysis of k-means and Enhanced K-means clustering algorithm for data mining", *Int. J. Sci. Eng. Res.* Vol.3, (2012). https://pdfs.semanticscholar.org/c752/009f6372e89aa1f1417857b671b242a58854.pdf.

[7] H. Guruler, "A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with K-Mean", *Neural Comput. Appl.* (2016). http://doi:10.1007/s00521-015-2142-2.

[8] S. a Yang, J. Yoon, K. Kim, Y. Park, "Measurements of Morphological and Biophysical Alterations in Individual Neuron Cells Associated with Early Neurotoxic Effects in Parkinson ' s Disease", *Int. Soc. Adv. Cytom.* (2017) pp:510–518. http://doi:10.1002/cyto.a.23110.

[9] H. Jung, K. Chung, "Knowledge-based dietary nutrition recommendation for obese management", *Springer Sci.* (2016) pp:29–42. http://doi:10.1007/s10799-015-0218-4.

[10] A. Alsayat, H. El-Sayed, Efficient genetic K-Means clustering for health care knowledge discovery, in: 2016 IEEE/ACIS 14th Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2016, 2016: pp. 45–52. http://doi:10.1109/SERA.2016.7516127.

[11] Y. Li, C. Tao, Y. Tan, K. Shang, J. Tian, "Unsupervised Multilayer Feature Learning for Satellite Image Scene Classification", *IEEE Geosci. Remote Sens. Lett.* Vol.13, (2016) pp:157–161. http://doi:10.1109/LGRS.2015.2503142.

[12] N.M. Salem, Segmentation of white blood cells from microscopic images using K-means clustering, in: Natl. Radio Sci. Conf. NRSC, Proc., 2014: pp. 371–376. http://doi:10.1109/NRSC.2014.6835098.

[13] M.M. K. Date, "Brain Image Segmentation Algorithm using K-Means Clustering", *Int. J. Comput. Sci. Appl.* Vol.6, (2013) pp:285–289.

[14] R.D. Prasad, K.B.V.K. Sai, R.K. Sai, B.V. Manoj, "Content Based Image Retrieval using Color and Texture", *Signal Image Process. An Int. J.* Vol.3, (2012) pp:39–57. http://doi:10.5121/sipij.2012.3104.

[15] H. Meng Han, Q.L. Dong, L.B. Lin, P.K. Malakar, "The Potensial

of Double K-Means clusteringfor Banana Image Segmentation", *J. Food Process Eng.* Vol.37, (2014) pp:10–18. http://doi:10.1111/jfpe.12054.

[16] Z. Qiu-yu, L. Jun-chi, Z. Mo-yi, D. Hong-xiang, L. Lu, "Hand Gesture Segmentation Method Based on YCbCr Color Space and K- Hand Gesture Segmentation Method Based on YCbCr Color Space and K-Means Clustering", *Int. J. Signal Process. Image Process. Pattern Recognit.* Vol.8, (2015) pp:105–116. http://doi:10.14257/ijsip.2015.8.5.11.

[17] S.R. Dubey, P. Dixit, N. Singh, J.P. Gupta, "Infected Fruit Part Detection using K-Means Clustering Segmentation Technique", *Int. J. Artif. Intell. Interact. Multimed.* Vol.2, (2013) pp:65–72. http://doi:10.9781/ijimai.2013.229.

[18] G. Wang, Y. Zhao, J. Huang, Q. Duan, J. Li, "A K-means-based network partition algorithm for controller placement in software defined network", *2016 IEEE Int. Conf. Commun. ICC 2016*. (2016). http://doi:10.1109/ICC.2016.7511441.

[19] B.F. Solaiman, A. Sheta, "Energy optimization in wireless sensor networks using a hybrid K-means PSO clustering algorithm", *Turkish J. Electr. Eng. Comput. Sci.* Vol.24, (2016) pp:2679–2695. http://doi:10.3906/elk-1403-293.

[20] R. Campagni, D. Merlini, R. Sprugnoli, M.C. Verri, "Data mining models for student careers", *Expert Syst. Appl.* Vol.42, (2015) pp:5508–5521. http://doi:10.1016/j.eswa.2015.02.052.

[21] D. Kabakchieva, "Predicting student performance by using data mining methods for classification", *Cybern. Inf. Technol.* Vol.13, (2013) pp:61–72. http://doi:10.2478/cait-2013-0006.

[22] H.I. Arumawadu, Rathnayaka, R M Kapila Tharanga, S.K. Illangarathne, "Mining Profitability of Telecommunication Customers Using K-Means Clustering", *J. Data Anal. Inf. Process.* Vol.3, (2015) pp:63–71. http://doi:10.4236/jdaip.2015.33008.

[23] B.A. Tama, "Data Mining For Predicting Customer Satisfaction in Fast-Food Restaurant", *J. Theor. Appl. Inf. Technol.* Vol.75, (2015) pp:18–24.

[24] F. An, H.J. Mattausch, "K-means clustering algorithm for multimedia applications with flexible HW/SW co-design", *J. Syst. Archit.* Vol.59, (2013) pp:155–164. http://doi:10.1016/j.sysarc.2012.11.004.

[25] M. Ay, O. Kisi, "Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques", *J. Hydrol.* Vol.511, (2014) pp:279–289. http://doi:10.1016/j.jhydrol.2014.01.054.

[26] S. Kumar, D. Toshniwal, "A data mining approach to characterize road accident locations", *J. Mod. Transp.* Vol.24, (2016) pp:62–72. http://doi:10.1007/s40534-016-0095-5.

[27] B. Choubin, K. Solaimani, M. Habibnejad Roshan, A. Malekian, "Watershed classification by remote sensing indices: A fuzzy c-means clustering approach", *J. Mt. Sci.* Vol.14, (2017) pp:2053–2063. http://doi:10.1007/s11629-017-4357-4.

[28] M. Yesilbudak, Clustering analysis of multidimensional wind speed data using k-means approach, in: Int. Conf. Renew. Energy Res. Appl., 2016: pp. 961–965. http://doi:10.1109/ICRERA.2016.7884477.

[29] B. Liu, G. Xu, Q. Xu, N. Zhang, "Outlier Detection Data Mining of Tax Based on Cluster", *Phys. Procedia*. Vol.33, (2012) pp:1689–1694. http://doi:10.1016/j.phpro.2012.05.272.

[30] N.R. Ravi, K. Vani, D. Gupta, "Exploration of Fuzzy C Means Clustering Algorithm in External Plagiarism Detection System", *Adv. Intell. Syst. Comput.* Vol.384, (2016) pp:127–128. http://doi:10.1007/978-3-319-23036-8.

[31] M. Kim, D.K. Han, H. Ko, "Joint patch clustering-based dictionary learning for multimodal image fusion", *Inf. Fusion*. (2015) pp:34–36. http://doi:10.1016/j.inffus.2015.03.003.

[32] J. Agarwal, R. Nagpal, R. Sehgal, "Crime Analysis using K-Means Clustering", *Int. J. Comput. Appl.* Vol.83, (2013) pp:1–4.

[33] T. Velmurugan, "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data", *Appl. Soft Comput. J.* Vol.19, (2014) pp:134–146. http://doi:10.1016/j.asoc.2014.02.011.

[34] P. Puri, I. Sharma, "Enhancement in K-mean Clustering to Analyze Software Architecture Using Normalization", *Int. J. Sci. Eng. Res.* Vol.6, (2015) pp:604–611.

[35] M. Silic, G. Delac, S. Srbljic, "Prediction of atomic web services reliability based on k-means clustering", *Proc. 2013 9th Jt. Meet. Found. Softw. Eng.* (2013). http://doi:10.1145/2491411.2491424.