Improving Clustering Method Performance Using K-Means, Mini batch K-Means, BIRCH and Spectral

Tenia Wahyuningrum
Department of Informatics
Institut Teknologi Telkom Purwokerto
Banyumas, Indonesia
tenia@ittelkom-pwt.ac.id

Siti Khomsah
Department of Data Science
Institut Teknologi Telkom Purwokerto
Banyumas, Indonesia
siti@ittelkom-pwt.ac.id

Suyanto Suyanto
School of Computing
Telkom University
Bandung, Indonesia
suyanto@telkomuniversity.ac.id

Selly Meliana
School of Computing
Telkom University
Bandung, Indonesia
sellym@telkomuniversity.ac.id

Prasti Eko Yunanto
School of Computing
Telkom University
Bandung, Indonesia
gppras@telkomuniversity.ac.id

Wikky F. Al Maki
School of Computing
Telkom University
Bandung, Indonesia
wikkyfawwaz@telkomuniversity.ac.id

***Abstract*— The most pressing problem of the *k*-Nearest Neighbor (KNN) classification method is voting technology, which will lead to poor accuracy of some randomly distributed complex data sets. To overcome the weaknesses of KNN, we developed a new scheme in data set clustering, making the number of clusters greater than the number of data classes. In addition, the committee selects each cluster so that it does not use voting techniques such as standard KNN methods. This study uses two sequential methods, namely the clustering method and the KNN method. Clustering methods can be used to group records into multiple clusters to select commissions from these clusters. Five clustering methods were tested: K-Means, K-Means with Principal Component Analysis (PCA), Mini Batch K-Means, Spectral and Balanced Iterative Reduction, and Clustering using Hierarchies (BIRCH). All tested clustering methods are based on the cluster type of the center of gravity. According to the result, the BIRCH method has the lowest failure rate among the five clustering methods (2.13), and K-Means has the largest clusters (156.63).**

Keywords—clustering, KNN, K-Means, BIRCH, Spectral

I. INTRODUCTION

The *k*-Nearest Neighbor (KNN) method or also known as the *k*-Nearest Neighbor Rule (KNNR) is a non-parametric classification method that is known to be the simplest, effective, good performance, and robust [1], [2]. This method works by finding a number of *k* patterns (among all the training patterns in all classes) closest to the input pattern, then determining the decision class based on using voting technique. Some of the weaknesses of the KNN method are that it is sensitive to less relevant features and the neighboring size of *k* [3], [4]. It is relatively difficult to determine the exact k because it can be high; in other cases, it can be very low. The most urgent problem in KNN is the voting technique, which makes it low-accuracy for several complex datasets which are randomly distributed [5]. To overcome the weakness of KNN, we created a new scheme in the form of dataset clustering so that the number of clusters is greater than the number of data classes. Furthermore, commissions will select each cluster, so it does not use voting techniques like the standard KNN method.

Clustering is a method of grouping data. According to [6] clustering is a process for grouping data into several clusters or groups so that the data in one cluster has the maximum level of similarity and the data between clusters has the minimum similarity. Five clustering methods are tested: K-Means, K-Means with Principal Component Analysis (PCA), Mini Batch K-Means, Spectral, and Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). All the clustering method tested is a type of centroid based clustering. The K-Means method is the most common clustering method [7], [8]. K-Means can group large amounts of data with fast and efficient computation time. K-Means is one of the clustering algorithms with a partitioning method based on the center point (centroid) [8]–[10]. Several studies use a combination of K-Means and PCA methods to reduce data dimensions in optimization clusters [11]. PCA is a technique used to simplify data by transforming the data linearly to form a new coordinate system with maximum variance [12]. The reason for using PCA is that reducing dimensions can eliminate irrelevant features, and reduce noise, while also reducing the curse of dimensionality [13].

Mini Batch K-Means is a modification of the K-Means algorithm to reduce computation time. The Mini Batch is part of the input data, randomly sampled at each training iteration, drastically reducing the amount of computation [14]. Mini Batch K-Means concentrates advantages faster than K-Means, but the quality is slightly reduced than K-Means [14], [15]. In Spectral Clustering, grouping is based on the similarity between each data. The similarity is seen in the relationship between each data; the clustering formed a graph from the existing data. The edge is a relationship between data which is usually the distance between two related records [16], [17]. Spectral clustering is considered a popular and effective method, but it fails to consider higher-order structures and performs poorly on directed networks [18]. The BIRCH algorithm is a hierarchical clustering method to find good clustering with just one data scan. The clustering quality improved with a few additional scans[19].

This study uses two consecutive methods, namely the clustering method and the KNN method. The clustering method is useful for grouping datasets into several clusters to

select commissions from these clusters. The KNN method is used for the classification process with training data from the comned commissions. We use a public dataset that has been tested with 15 datasets. This research is a small part of improving the performance of KNN classification by applying a data commission (not using all data as training data). The selection of this commission is carried out by a clustering process first. Then, a commission representative will be selected for the KNN process from the members of each resulting cluster.

We assume that if a more efficient clustering method is chosen, the clusters formed will be more homogeneous. Improving the clustering method is expected to produce more homogeneous clusters that will provide commissions that are more representative of the cluster. As a result, the performance of the KNN classification will increase. A measuring tool for KNN performance is the error rate in each classification process. Each method is also tested for the best K value (number of clusters) for each test dataset.

II. Research Method

A. Research method and experiment scenario

The proposed method is to determine the dataset, then perform clustering based on the K-Means, Minibatch K-Means, BIRCH, and Spectral methods. Each clustering method will be tested to find the best value of K (number of clusters) with silhouette testing. The next step is the selection of commissions on each realized cluster and further classification is carried out using the KNN method (Figure 1).

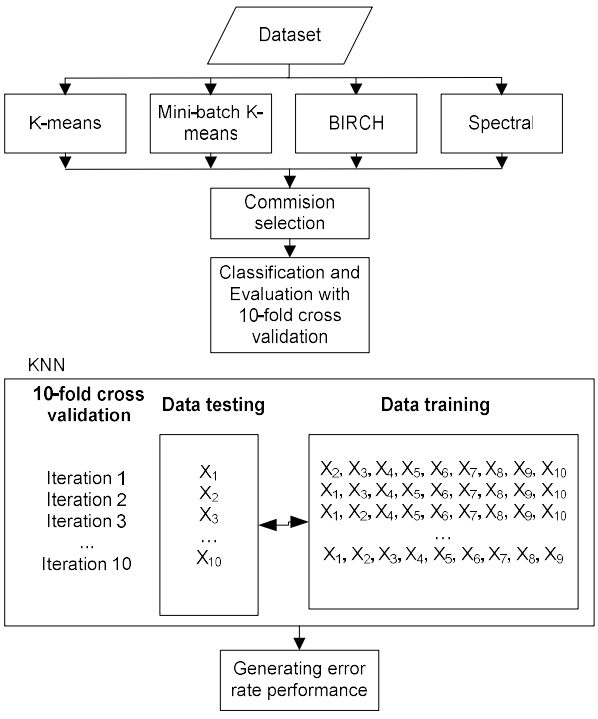

Fig. 1. Experiment scenario

B. K-Means

K-means is a classical clustering algorithm that is widely used in various fields. The basic principle of the K-means algorithm is to calculate the distance of each data point from the center of the cluster. Then the average value of each cluster will be measured with a simple algorithm, then divide the class iteratively until all clusters are covered. This process is repeated until it converges. At present, the K-means algorithm has become one of the important methods of machine learning and data mining technology[10].

C. Mini Batch K-Means

The K-Means clustering method was upgraded to a method called Mini Batch K-Means. Different from K-Means, this one does not use all data records in the dataset every time the clustering iteration process is carried out, but chooses a subset of records randomly from the dataset for distance calculations. This greatly reduces clustering time, and overall reduces convergence time [20].

D. Spectral Clustering

Spectral Clustering is a grouping method based on similarities between one data and another. The similarity is seen from the relationship between each data. Spectral Clustering will form a graph from existing data where the vertices of the graph are data and the edges are relationships between data which are usually the distance between two related records. Spectral clustering is one of the simplest clustering algorithms to implement, can be solved efficiently with standard linear algebra software, and very often outperforms traditional clustering algorithms such as the k-means algorithm. Eventually spectral clustering became one of the most popular clustering algorithms [21].

E. BIRCH

The BIRCH algorithm is suitable for dealing with the problem of grouping discrete and continuous attribute data. BIRCH applies an integrated hierarchical principle using feature clustering (CF) and feature tree clusters (CF Tree). The clustering feature tree describes useful clustering of information, and takes advantage of much less memory space than the size of data that can be stored in memory. So this algorithm can improve performance in clustering large data sets at high speed and wide scalability [22].

F. PCA

PCA is a procedure used in several fields, such as face recognition and image recognition[13]. PCA is a technique used to reduce attributes with a very small risk of losing information [23], [24]. In addition, the PCA technique produces a specific subspace for each cluster and is useful for better clustering processes [25]. PCA changes the variables that were n variables to be reduced to k new variables (principal components) with the number of k less than n. PCA also allows only using k principal components to produce the same value using n variables.

G. Commision Selection

Commissions are generated and selected from the formed clusters. The percentage of commission taken from each cluster is determined by trial, but not more than 50%. These commissions will be part of the nearest data search process on the k-NN algorithm and the results of this closest data search will be tested using several standard testing methods.

H. Classification and Validation

Classification is done by dividing the commission dataset into training data and testing data with a ratio of 80:20. The training data is used to form a classification model, and the testing data will be used to test the model. Validation using k-fold cross validation, the dataset will be divided into 10 folds, and in turn one fold dataset will function as testing data, and another 9 folds as training data. Then the experiment will occur 10 times. Each experiment will get an error rate value. The error rate value shows the performance of the classification model. The application of 10-fold cross validation produces 10 error rate values. So we determined that the reported error rate is the average error rate of each trial in the 10-fold cross validation. Cluster number is the average cluster obtained from each experiment. We use the number of clusters as a measure of the success of the clustering process. We assume that the more clusters formed, the more homogeneous the commissions sent for the KNN process. However, the drawback is that the process is complex. Especially for K-Means we added a Principal Component Analysis (PCA) process. PCA as is well known is useful for simplifying data. It is hoped that the K-Means process will improve its performance better, and can be compared with the BIRCH, Spectral and Mini Batch K-Means methods. The specific flowchart for the K-Means process with PCA is shown in Figure 2.

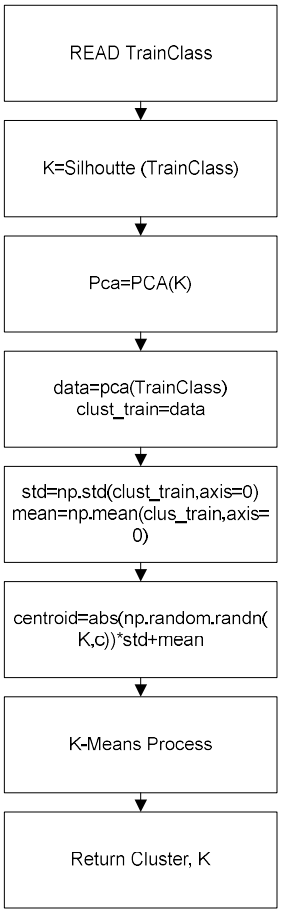

Fig. 2. K-Means model and PCA

I. Dataset

Fifteen of the datasets that we have collected from University of California Irvine (UCI) Machine Learning and other repository are in Table I.

TABLE I. DATASETS USED IN THIS RESEARCH

No	Dataset	Record Number	Column Number	Class Number
1	Sonar	208	61	2
2	Seed	210	8	3
3	Wine	178	14	3
4	Bank	1372	5	2
5	Iris	150	6	3
6	CNAE	1080	857	9
7	Pima Indian	768	9	2
8	Park	195	23	2
9	Libras	360	91	15
10	Climate	540	21	2
11	Plrx	182	13	2
12	QSAR	1055	42	2
13	Ecoli	307	9	4
14	Haber	306	4	2
15	Musk1	476	167	2

Each dataset varies the number of records, the number of classes and the number of columns. The characteristics of the dataset will affect the classification process. We also want to know the relationship between the number of records, the number of columns and the class number with the error rate in each clustering test in this study.

III. Result And Discussion

A. Research Result

This research process will calculate the classification error rate and calculate the number of clusters for the dataset segmentation process so that it can be continued in the selection of commissions for the KNN process. So every time the test will be recorded error rate and number of clusters. Experiments were carried out on each dataset. The commission selection process for KNN will be tested for the clustering process using K-Means, K-Means with PCA, BIRCH, Spectral Clustering and Mini Batch Clustering.

The error rate is usually used in clustering problems to see what percentage of predictions are wrong, while accuracy calculates the percentage of correct predictions. The accuracy value ranges from 0 to 1 or equivalent to a percentage of 0 to 100%, meaning that the higher the accuracy value, the better an algorithm will be. The accuracy and error rate are complements of each other, meaning that we can always calculate one from the other.

$$error\ rate = 100\% - \left(\left(\frac{a}{b}\right) \times 100\%\right) \quad (1)$$

where a is number of correct predictions, b is number of data.

$$accuracy = 100\% - error\ rate \quad (2)$$

Therefore, we focused on analyzing the error rate as one of the classification performance indicators at the model development stage. Further analysis, to find out whether there is a significant difference in the clustering method with the lowest error rate, multivariable mean analysis is used. The results of the error rate and number of clusters are summarized in Table II.

TABLE II. RESEARCH RESULT

No	Dataset	Cluster Number				Average Error Rate (%)				
		K-Means	Mini Batch K-Means	BIRCH	Spectral	K-Means	K-Means PCA	Mini Batch K-Means	BIRCH	Spectral
1	Sonar	15	15	19	5	45.76	50.38	38.53	**37.05**	38.00
2	Seed	21	18	12	12	37.14	29.98	**10.48**	**10.48**	11.90
3	Wine	9	9	10	6	28.53	**25.65**	25.75	26.31	25.75
4	Bank	4	4	4	5	4.53	**0.07**	**0.07**	**0.07**	**0.07**
5	Iris	6	6	6	6	7.33	**4.00**	**4.00**	**4.00**	**4.00**
6	CNAE	34	31	47	29	16.95	11.02	**9.08**	10.46	10.09
7	Pima Indian	4	4	4	4	27.47	26.45	26.71	26.71	**26.32**
8	Park	7	8	6	7	23.61	23.11	**20.66**	22.24	22.68
9	Libras	75	74	55	75	44.44	17.50	18.06	17.22	**16.94**
10	climate	13	11	14	4	**8.89**	9.45	9.08	8.9	11.85
11	Plrx	20	17	14	8	54.65	47.87	**29.09**	31.87	43.98
12	QSAR	7	9	9	4	21.90	19.54	**19.26**	**19.26**	19.83
13	Ecoli	9	9	9	10	5.26	5.58	3.97	**3.30**	3.97
14	Haber	6	5	6	5	41.77	36.52	35.17	34.56	**33.26**
15	Musk1	6	6	5	4	20.31	17.39	**17.18**	17.81	18.44

B. *Number Cluster*

The highest number of clusters produced is the K-Means method and the least number of clusters is the Spectral method. Fig. 3 shows that the minimal cluster method is dominated by the Spectral clustering method on the Sonar, Seed, Wine, Bank, Iris, CNAE, Pima Indian, Parkinson, Libras, climate, Plrx, QSAR, Ecoli, Haber, Musk1 datasets. Although the average number of clusters in the K-Means method is the largest, the BIRCH method more often dominates the maximum number of clusters. It can be seen in Fig. 3 the maximum number of clusters produced by the BIRCH method is in the Sonar, Wine, CNAE, Pima Indian, Climate, and Haber datasets.

C. *Error Rate Performance*

The Friedman test was carried out with observations in each group that were ranked separately so that it had a ranking data cluster with five treatments. Table III shows the Friedman test ranks.

TABLE III. FRIEDMAN TEST RANKS

Method	Mean Rank
K-Means without PCA	4.63
K-Means with PCA	3.33
MiniBatch-Kmeans	2.30
BIRCH	2.13
Spectral	2.60

Table III explains the output ranks table shows the average error rate in ranking form. For example, the highest average error rate is in the K-Means method without PCA, and the lowest error rate is the BIRCH method. This value shows that the average error rate in BIRCH is the smallest, or it can be said that the accuracy of the BIRCH clustering method is quite high.

H_0: There is no significant difference in error rate performance between the five clustering methods.

H_1: There is a significant difference in error rate performance between the five clustering methods.

Guidance in decision-making can be seen from the asymptotic significance value using significance level 5%. Therefore, if the asymptotic significance value is more than 0.05, then H_0 is accepted, and vice versa. Based on Table IV, it can be seen that the asymptotic significance value is 0.000 <0.05. Then H_0 is rejected, or it can be said that there is a difference in the average error rate between the five methods.

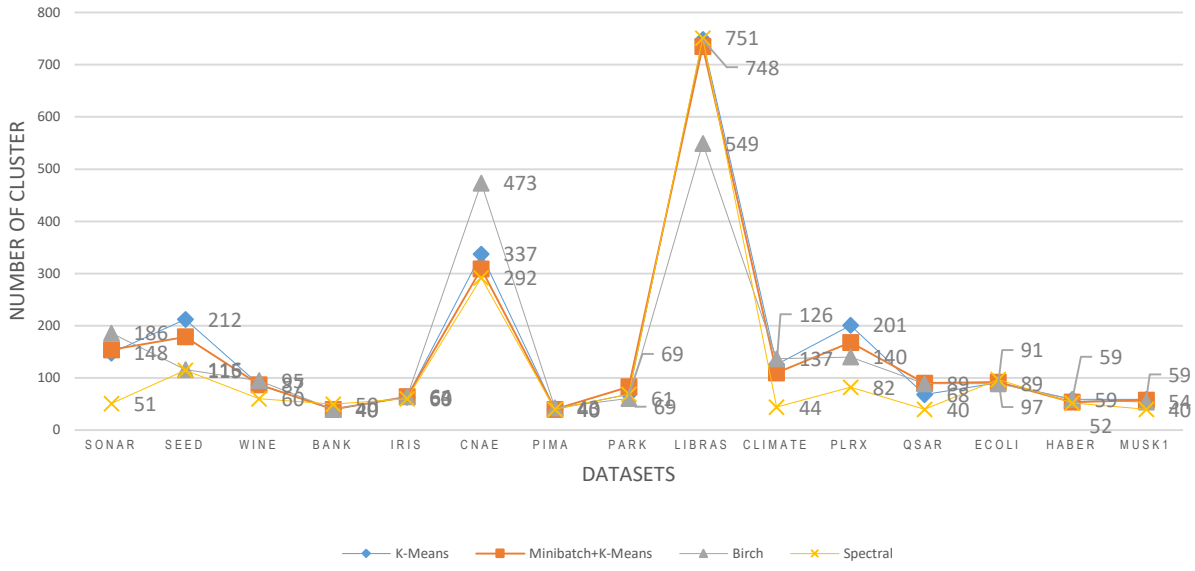

Fig. 3. Number of Cluster

TABLE IV. TEST STATISTICS

N	15
Chi-Square	27.261
df	4
Asymp. Sig.	0.000
[a.] Friedman Test	

The other way to accepted or rejected a hypothesis is by chi-square value. According to Table IV, the calculated chi-square value of 27.261 is larger than the value of the chi-square table with a degree of freedom (df) 4, which is 9.488. So, the conclusion is H_0 rejected.

IV. CONCLUSION

Based on the test result and discussion, the lowest error rate performance is the BIRCH method of the five clustering methods, while the one with the highest number of clusters is K-Means. So this study proposes improvement of the KNN method using the BIRCH method for the clustering process in the selection of commissions so that the performance of KNN is more accurate. For future work, the classification will be performed by experimentally comparing training data, and test data at 50:50, 60:40, with the number of rows of the data sets being increased to approximately 10,000.

REFERENCES

[1] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A New K-Nearest Neighbors Classifier for Big Data Based on E ffi cient Data Pruning," pp. 1–12.

[2] M. M. Kumbure, P. Luukka, and M. Collan, "A new fuzzy k -nearest neighbor classifier based on the Bonferroni mean," *Pattern Recognit. Lett.*, vol. 140, pp. 172–178, 2020, doi: 10.1016/j.patrec.2020.10.005.

[3] N. Rastin, M. Z. Jahromi, and M. Taheri, "A Generalized Weighted Distance k-Nearest Neighbor for Multi-label Problems," *Pattern Recognit.*, vol. 114, p. 107526, 2021, doi: 10.1016/j.patcog.2020.107526.

[4] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of Improving K-Nearest-Neighbor for Classification," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2007, pp. 679–683.

[5] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006, doi: 10.1007/s10462-007-9052-3.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edit. Waltham, USA, 2012.

[7] H. V. Bhagat and M. Singh, "A Comparative Approach to Evaluate Different CVIs using Grid K-Means and Improved K-Means Clustering," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, 2020.

[8] J. Xu, D. Han, K. Li, and H. Jiang, "A K-means algorithm based on characteristics of density," *Comput. Sci. Inf. Syst.*, vol. 17, no. 2, pp. 665–687, 2020.

[9] A. H. Khaleel and I. Q. Abduljaleel, "A novel technique for speech encryption based on k-means clustering and quantum chaotic map," *Bull. Electr. Eng. Informatics*, vol. 10, no. 1, pp. 160–170, 2021, doi: 10.11591/eei.v10i1.2405.

[10] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 248–254, 2020, doi: 10.14569/IJACSA.2020.0110832.

[11] L. Cappelletti *et al.*, "Complex Data Imputation by Auto-Encoders and Convolutional Neural Networks — A Case Study on Genome Gap-Filling," doi: 10.3390/computers9020037.

[12] A. A. Miranda, Y. A. Le Borgne, and G. Bontempi, "New routes from minimal approximation error to principal components," *Neural Process. Lett.*, vol. 27, no. 3, pp. 197–207, 2008, doi: 10.1007/s11063-007-9069-2.

[13] L. Smith, "A tutorial on PCSA," University of Otago, New Zealand, 2006.

[14] A. Feizollah, N. B. Anuar, R. Salleh, and F. Amalina, "Comparative Study of K-means and Mini Batch K-means Clustering Algorithms in Android Malware Detection Using Network Traffic Analysis," in *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, 2014, no. February, pp. 193–197.

[15] M. Arunachalamand and K. Amuthan, "Finger Knuckle Print Recognition using MMDA with Fuzzy Vault," *Int. Arab J. Inf. Technol.*, vol. 17, no. 4, pp. 554–561, 2020.

[16] S. Trivedi, Z. A. Pardos, G. N. Sarkozy, and N. T. Heffernan, "Spectral Clustering in Educational Data Mining," in *4th International Conference on Educational Data Mining*, 2011, no. January, pp. 129–138.

[17] X. Cai and F. Sun, "Supervised and Constrained Nonnegative Matrix Factorization with Sparseness for Image Representation," *Wirel. Pers. Commun.*, vol. 102, no. 4, pp. 3055–3066, 2018, doi: 10.1007/s11277-018-5325-1.

[18] W. G. Underwood, A. Elliot, and M. Cucuringu, "Motif-based spectral clustering of weighted directed networks," *Appl. Netw. Sci.*, vol. 5, no. 62, pp. 1–41, 2020.

[19] B. Denclue and M. C. Nwadiugwu, "Gene-Based Clustering Algorithms : Comparison," *Bioinform. Biol. Insights*, vol. 14, pp. 1–6, 2020, doi: 10.1177/1177932220909851.

[20] K. Peng, V. C. M. Leung, and Q. Huang, "Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data," *IEEE Access*, vol. 6, no. March, pp. 11897–11906, 2018, doi: 10.1109/ACCESS.2018.2810267.

[21] U. von Luxburg, "A Tutorial on Spectral Clustering," 2007.

[22] F. Ramadhani, M. Zarlis, and S. Suwilo, "Improve BIRCH algorithm for big data clustering," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 725, no. 1, 2020, doi: 10.1088/1757-899X/725/1/012090.

[23] C. C. Aggrawal, *Data Mining*. New York, USA: Springer, 2015.

[24] S. Land, *RapidMiner Operator Reference Manual*. 2014.

[25] C. C. Aggrawal and C. Zhai, *Mining Text Data*. New York, USA: Springer, 2012.