STMIK AKAKOM YOGYAKARTA
Code for Life

4th ISRITI 2021

Improving Clustering Method Performance using K-Means, Mini Batch K-Means, BIRCH and Spectral

Tenia Wahyuningrum, Siti Khomsah, Suyanto, Selly Meliana,
Prasti Eko Yunanto, Wikky F. Al Maki

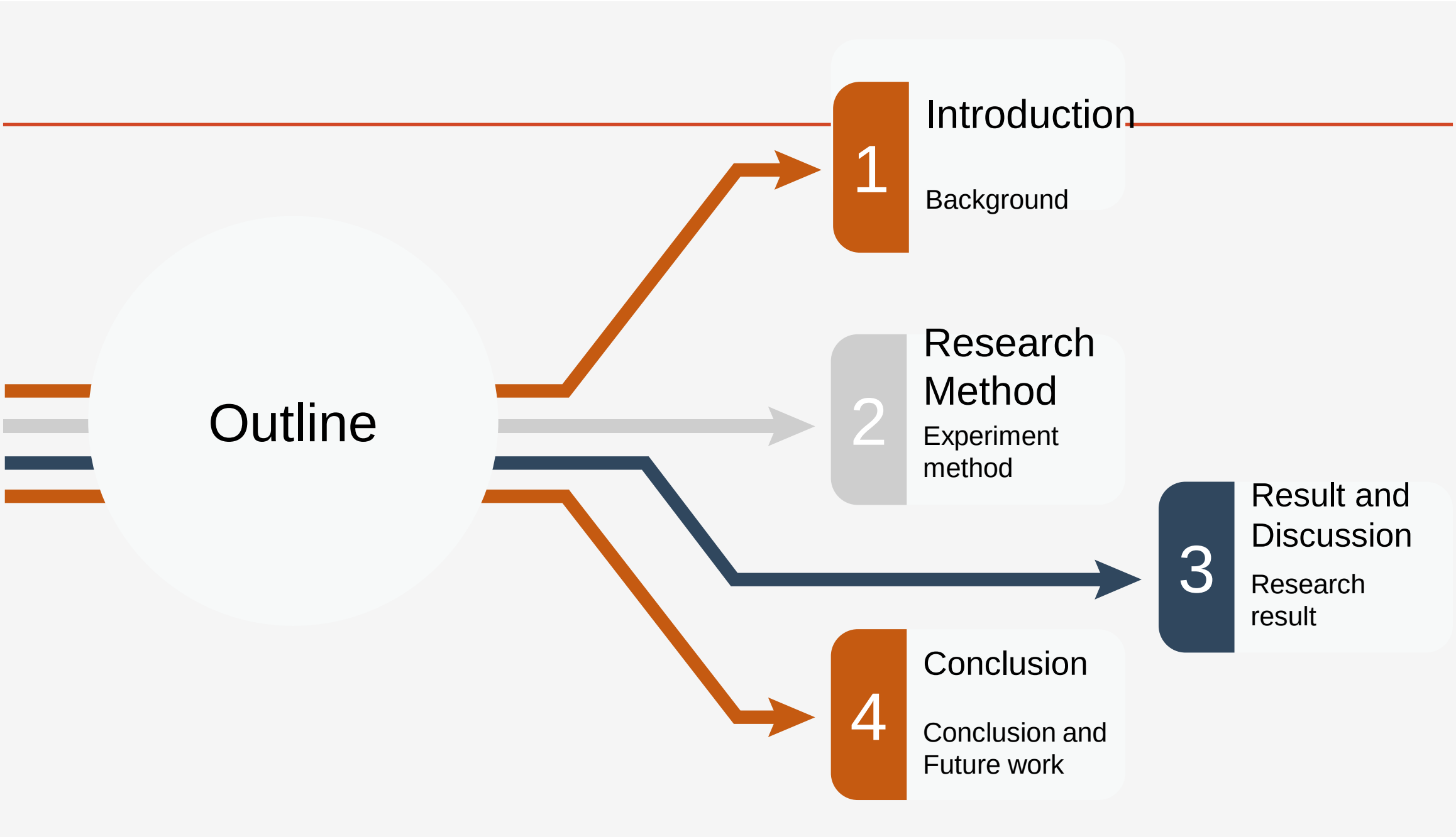
tenia@ittelkom-pwt.ac.id

IEEE Advancing Technology for Humanity

IEEE INDONESIA SECTION

IEEE STUDENT BRANCH STMIK AKAKOM

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA


Outline
1
Introduction
Background
2
Research Method
Experiment method
3
Result and Discussion
Research result
4
Conclusion
Conclusion and Future work

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**.

Introduction

- ✓ Clustering is a method of grouping data. According to Han, et al [6], clustering is a process for grouping data into several clusters or groups in one cluster has the maximum 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).*
- ✓ 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.

Research Method

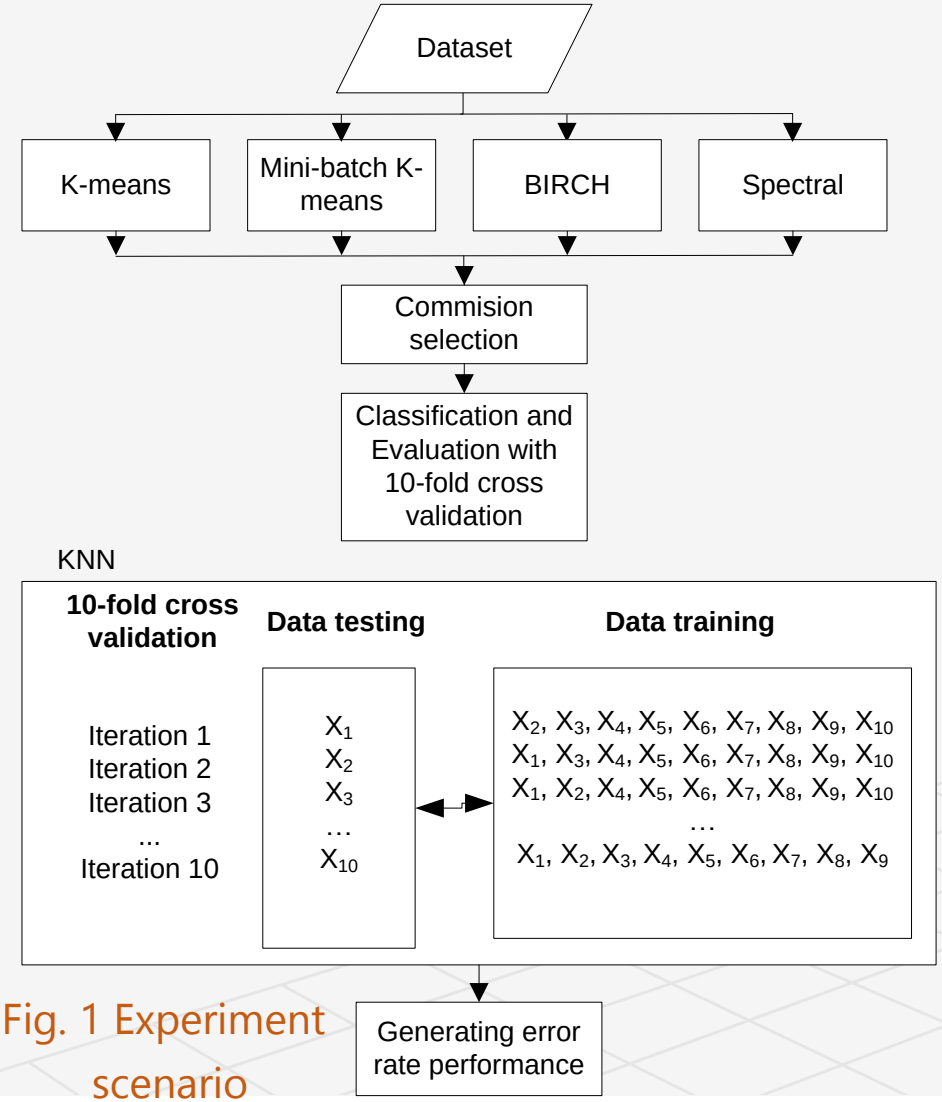


Fig. 1 Experiment scenario

The proposed method determines the dataset and then perform clustering based on the **K-Means, Minibatch K-Means, BIRCH, and Spectral methods**. First, 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.

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

Fifteen of the datasets we have collected from the University of California Irvine (UCI) Machine Learning and other repository

Result and Discussion

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

Research result

Based on this result, it can be seen that the lowest error rate value is 0.07 in the BANK dataset using the K-Means with PCA, Mini Batch K-Means, BIRCH, and Spectral methods. In comparison, the highest error rate value is 54.65, namely the Plrx (Planning Relax) dataset using K-Means.

Result and Discussion

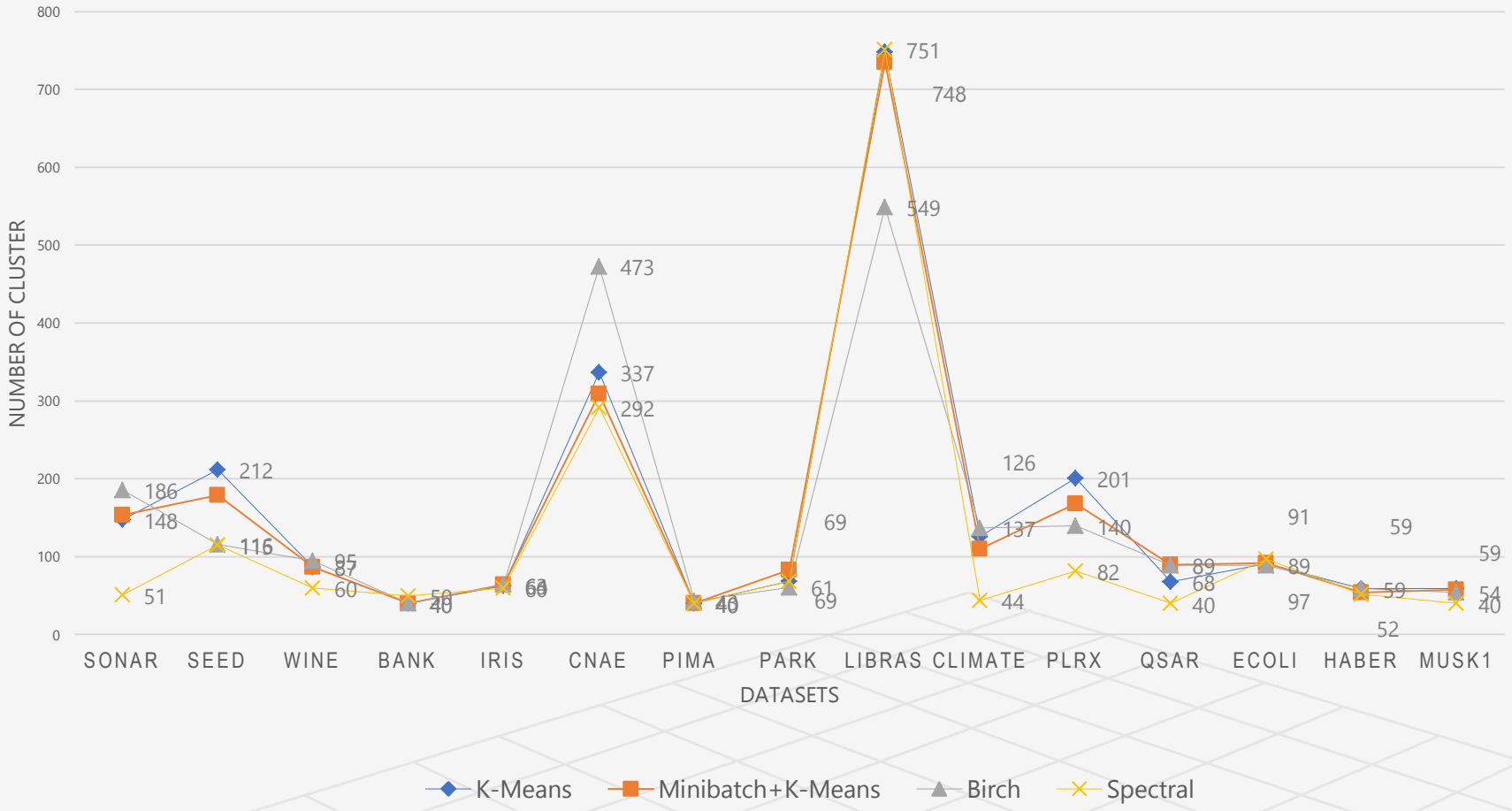


Fig. 3 Number cluster

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.

Result and Discussion

Friedman test

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

This table explains the output ranks; the table shows the average error rate in ranking form. 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.

Result and Discussion

Test statistics

N	15
Chi-Square	27.261
df	4
Asymp. Sig.	0.000

[a.] Friedman Test

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.

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.

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.

Thank You