Optimization of Unsupervised Learning in Machine Learning

Hidayatus Sibyan¹, Wildan Suharso², Edi Suharto³, Melda Agnes Manuhutu⁴, Agus Perdana Windarto⁵*

¹Universitas Sains Al-Qur'an Jawa Tengah, Indonesia
²Universitas Muhammadiyah Malang, Indonesia
³Politeknik Pikesi Ganesha Bandung, Indonesia
⁴Universitas Victory Sorong, Indonesia
⁵STIKOM Tunas Bangsa, Pematangsiantar, Indonesia

Email: *agus.perdana@amiktunasbangsa.ac.id

Abstract. The Ombudsman of the Republic of Indonesia (hereinafter referred to as the Ombudsman) is a state institution (independent) that has the authority to oversee the administration of public services. The purpose of this study is to analyze the completion of reports/complaints from the public by using unsupervised learning techniques in machine learning. The data source used is the statistical report/public complaints based on the classification of the reporter and how to submit it in each provincial regional office (simpel.ombudsman.go.id). The unsupervised learning techniques in machine learning that are used are clustering (k-medoids) and classification (C4.5) which are part of data mining. k-medoids is tasked with mapping community reports/complaints based on provincial regional offices. The results of the mapping will be classified to get the range of values from the existing mapping. The calculation process uses the help of RapidMiner software. The distribution labels used were 4 clusters namely the percentage of completion of the "very good" report (C1) of 9 provinces; percentage of "good" report completion (C2) of 10 provinces; percentage of completion of "lacking" reports (C3) of 11 provinces; percentage of "bad" report completion (C4) of 3 provinces. The Davies-Bouldin Index value on the map is 0.530 (optimal). The results of the mapping can be information in improving the quality of public services in the completion of the report including the provinces included in the C3 and C4 clusters with the percentage of report completion classification (C4 = 0 - 20.70% and 20.70% > C3 < 47.69%).

1. Introduction

The Ombudsman of the Republic of Indonesia (called the Ombudsman) is a state institution (independent) that has the authority to oversee the administration of public services. This is stated in article 2 of the Law of the Republic of Indonesia Number 37 of 2008 concerning the Ombudsman of the Republic of Indonesia and article 3 of the Law of the Republic of Indonesia Number 37 of 2008 concerning the Ombudsman in carrying out his duties and authorities. Since the establishment of the Ombudsman as a state institution, it has received many reports/complaints from the public which are presented in the form of statistics reports/complaints from the public through simpel.ombudsman.go.id. The number of reports/complaints from the public consists of several classifications such as how to submit, report status, group of reported institutions, alleged maladministration and report substance. Based on this database, the purpose of this study is to explore...
the information contained therein by utilizing unsupervised learning techniques in machine learning. The unsupervised learning technique is a method that is applied without requiring training data. Unsupervised learning makes observations or data without any label/class/decision. Some algorithms that can be used in unsupervised learning are k-means, k-medoids, hierarchical clustering, dbscan, fuzzy c-means and self-organizing map [1][2]–[7].

Data mining techniques are part of unsupervised learning consisting of clustering, classification, estimation, association and pattern recognition [3], [6], [8]–[10]. Data mining is a technique of extracting knowledge from a large amount of data (warehouse) which will find various types of patterns inherited in the data so that it is useful [11]. Clustering is one technique that can be used to do leveling [12], [13]. Popular methods are k-means and k-medoids [14], [15]. Both of these methods have their advantages. In addition, k-medoids is a development of k-means variants so that it appears as a remedy for weaknesses of k-means that are sensitive to outlier [14] and reduce the sensitivity of the resulting partition with respect to extreme values contained in the dataset [15], [16]. In addition, many previous studies using k-medoids in solving problems among [17] regarding the grouping of student scholarships. The greatest value of purity on the whole dataset of data codification is 91.67%, it can be concluded that the K-Medoids algorithm is more suitable for use in a dataset with encoded attributes. In addition, this study wants to combine the two methods of k-medoids (clustering) and C4.5 (classification) [18]–[20] in mapping community reports/complaints based on provincial representative offices. In this case the k-medoid method is tasked with mapping community reports/complaints based on the provincial regional office. The mapping results will be classified to get the range of values from the existing mapping. It is expected that the mapping results can be information in improving the quality of public services in the completion of reports/public complaints.

2. Methodology

Source of research data used is Ombudsman data through the url simpel.ombudsman.go.id. The data is the number of community reports based on the reporter's classification and method of delivery. The delivery method can be done by direct arrival, email, facsimile, investigation, initiative, media, letter, telephone and website (table 1). Table 1 is the total community reports based on the representative offices of each province and table 2 is the number of completed reports obtained from the ombudsman report of the Republic of Indonesia. The following is complete research data.

### Table 1. Number of Public Reports Based on Reporting Classifications and Methods of Submission

| Province            | Come Live | Email | Facsimile | Investigation Initiative | Media | Letter | Telephone | Website | Report Total |
|---------------------|-----------|-------|-----------|--------------------------|-------|--------|-----------|---------|--------------|
| Aceh                | 85        | 14    | 0         | 10                       | 38    | 12     | 16        | 0       | 175          |
| Bali                | 82        | 0     | 0         | 12                       | 68    | 11     | 17        | 1       | 191          |
| Banten              | 38        | 8     | 0         | 49                       | 20    | 5      | 0         | 0       | 120          |
| Bengkulu            | 88        | 2     | 0         | 6                        | 17    | 4      | 10        | 0       | 127          |
| DI Yogyakarta       | 114       | 3     | 0         | 8                        | 86    | 21     | 0         | 233     |              |
| DKI Jakarta         | 370       | 24    | 1         | 0                        | 0     | 726    | 1         | 0       | 1122         |
| Gorontalo           | 113       | 0     | 0         | 6                        | 3     | 3      | 11        | 0       | 136          |
| Jambi               | 57        | 1     | 0         | 5                        | 39    | 11     | 4         | 0       | 117          |
| West Java           | 76        | 6     | 0         | 7                        | 3     | 21     | 0         | 0       | 113          |
| Central Java        | 57        | 1     | 0         | 1                        | 2     | 64     | 1         | 0       | 126          |
| East Java           | 138       | 5     | 1         | 5                        | 0     | 202    | 1         | 0       | 352          |
| West Kalimantan     | 121       | 1     | 0         | 2                        | 41    | 45     | 4         | 0       | 214          |
| South Borneo        | 59        | 0     | 0         | 0                        | 37    | 17     | 4         | 0       | 117          |
| Central Kalimantan  | 52        | 3     | 0         | 2                        | 45    | 7      | 0         | 0       | 109          |
| East Kalimantan     | 26        | 3     | 0         | 3                        | 39    | 13     | 1         | 0       | 85           |
| Bangka Belitung Islands | 89   | 0    | 0         | 27                       | 0     | 3      | 1         | 0       | 120          |
| Riau islands        | 63        | 3     | 0         | 4                        | 35    | 4      | 1         | 0       | 110          |
| Lampung             | 30        | 1     | 0         | 24                       | 25    | 21     | 7         | 0       | 108          |
| Maluku              | 104       | 0     | 0         | 0                        | 18    | 12     | 1         | 0       | 135          |
| North Maluku        | 28        | 0     | 0         | 71                       | 0     | 5      | 0         | 0       | 104          |
| West Nusa Tenggara | 99        | 5     | 0         | 24                       | 5     | 17     | 8         | 0       | 158          |
| East Nusa Tenggara | 146       | 1     | 0         | 2                        | 22    | 67     | 266       | 0       | 504          |
| Papua               | 119       | 4     | 0         | 4                        | 1     | 12     | 8         | 0       | 148          |
| West Papua          | 40        | 0     | 0         | 21                       | 11    | 14     | 35        | 0       | 121          |
| Riau                | 134       | 4     | 2         | 0                        | 14    | 32     | 15        | 0       | 201          |
| West Sulawesi       | 110       | 5     | 0         | 20                       | 60    | 5      | 8         | 0       | 208          |
Table 2. Number of Community Reports Based on Office/Representative

| Province     | Number of Reports | Report Closed | Percentage (%) |
|--------------|------------------|---------------|----------------|
| Aceh         | 175              | 87            | 49.71          |
| Bali         | 191              | 145           | 75.92          |
| Banten       | 120              | 70            | 58.33          |
| Bengkulu     | 127              | 106           | 83.46          |
| DI Yogyakarta| 233              | 80            | 34.33          |
| DKI Jakarta  | 1122             | 169           | 15.06          |
| Gorontalo    | 136              | 68            | 50.00          |
| Jambi        | 117              | 73            | 62.39          |
| West Java    | 113              | 50            | 44.25          |
| Central Java | 126              | 8             | 6.35           |
| East Java    | 352              | 199           | 56.53          |
| West Kalimantan | 214         | 86            | 40.19          |
| South Borneo | 117              | 73            | 62.39          |
| Central Kalimantan | 109       | 85            | 77.98          |
| East Kalimantan | 85        | 71            | 83.53          |
| Bangka Belitung Islands | 120 | 104         | 86.67          |
| Riau islands | 110              | 32            | 29.09          |
| Lampung      | 108              | 71            | 65.74          |
| Maluku       | 135              | 40            | 29.09          |
| North Maluku | 104              | 71            | 68.27          |
| West Nusa Tenggara | 158     | 134           | 84.81          |
| East Nusa Tenggara | 504     | 268           | 53.17          |
| Papua        | 148              | 39            | 26.35          |
| West Papua   | 121              | 6             | 4.96           |
| Riau         | 201              | 113           | 56.22          |
| West Sulawesi| 208              | 95            | 45.67          |
| South Sulawesi| 275             | 108           | 39.27          |
| Central Sulawesi | 153       | 137           | 89.54          |
| Southeast Sulawesi | 145   | 63            | 43.45          |
| North Sulawesi| 456             | 370           | 81.14          |
| West Sumatra | 271              | 95            | 35.06          |
| South Sumatra| 125              | 67            | 53.60          |
| North Sumatra| 180              | 72            | 40.00          |

Source: simpel.ombudsman.go.id

Figure 1. Research Methodology

The research methodology stage (fig.1), studying the research based on reliable sources is the initial stage that is a very important part of a research. Research data obtained through the source simpel.ombudsman.go.id will still be analyzed using RapidMiner software to clean up missing data. Because the data is one of the determinants of the results of research being carried out. After preprocessing the data, the data is processed using the k-medoids (clustering) method to see the mapping results. The results of the mapping are reprocessed using C.45 (classification) to see the
value of the rules of the cluster created. Then the results are tested with several parameters to see the results of the existing classification and classification. After that an analysis is taken based on the results of tests conducted to obtain conclusions.

3. Results and Discussion

Analysis of unsupervised learning techniques in machine learning for cases of public complaints to the ombudsman of the Republic of Indonesia through a representative office was carried out using the help of RapidMiner software. As explained in Figure 1, the analysis process uses a combination of k-medoids and C4.5. Cluster labels that are used in mapping the percentage of community complaints report completion are four clusters (the percentage of "very good" report completion (C1); the percentage of "good" report completion (C2); the percentage of "less" report completion (C3) and the percentage of report completion "bad"). Determination of the number of clusters (k) is done by comparing the values (k = 2; k = 3 and k = 4) where k = 4 has an optimal cluster value = 0.530 (Davies-Bouldin Index). Following is a picture of the k-medoids and C4.5 models along with the cluster results using the help of RapidMiner software.

Based on the calculation of the k-medoids method with the help of RapidMiner software, the results of four clusters are obtained, namely the percentage of "very good" report completion (C1) of 9 provinces; percentage of "good" report completion (C2) of 10 provinces; the percentage of completion of "poor" reports (C3) of 11 provinces and the percentage of completion of "bad" reports of 3 provinces). In this case the C4 cluster (cluster_0) consisting of DKI Jakarta, Central Java and West Papua is the "worst" cluster in completing the public complaint report. Following table centroid final results and mapping results using the k-medoids method.

| Attribute | C1 (cluster_2) | C2 (cluster_1) | C3 (cluster_3) | C4 (cluster_0) |
|-----------|----------------|----------------|----------------|----------------|
| Percentage (%) | 81,140 | 53,600 | 40 | 4,959 |
Based on the results of mapping using k-medoids, a classification result (C4.5) is obtained for the range of values from the cluster. The results obtained from cluster C1 (cluster_2) are ≥ 67,005; cluster C2 (cluster_1) is 47,695 - 67,005; cluster C3 (cluster_3) is 20,708 - 47,694; cluster C4 (cluster_0) is 0 - 20,707. The results of the classification can be seen in the following figure:

Based on Figure 6 the DBI results of 0.530 illustrate that the clustering of the percentage of community complaint reports based on provincial regional offices was formed quite well. Because essentially DBI wants the smallest value to assess the good cluster obtained. This value indicates the quality of membership in a cluster (intra-cluster) which has a high level of similarity and the distance of dissimilarity between clusters (inter-cluster) which is also quite high [21].
4. Conclusion

The results of research conducted using a combination of clustering and classification methods on unsupervised machine learning techniques can be applied properly. The results obtained from the calculation of the method that there are 3 provinces that are included in the category of "poor" in serving the completion of public complaints reports.

References

[1] A. Manconi and S. Loew, “Mapping Landslides on EO Data : Performance of Deep Learning Models vs . Traditional Machine Learning Models,” 2020.
[2] W. M. Sari et al., “Improving the Quality of Management with the Concept of Decision Support Systems in Determining Factors for Choosing a Cafe based on Consumers,” J. Phys. Conf. Ser., vol. 1471, no. 1, 2020.
[3] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, “Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination C,” 2020.
[4] H. Pratiwi et al., “Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks,” J. Phys. Conf. Ser., vol. 1471, no. 1, 2020.
[5] I. G. I. Sudipa, C. Astria, K. F. Irmanda, and A. Perdana, “Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses . Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses .” 2020.
[6] A. P. Windarto et al., “Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province,” J. Phys. Conf. Ser., vol. 1255, no. 1, 2019.
[7] R. H. S. Siburian, R. Karolina, P. T. Nguyen, E. L. Lydia, and K. Shankar, “Leaf disease classification using advanced SVM algorithm,” Int. J. Eng. Adv. Technol., vol. 8, no. 6 Special Issue, pp. 712–718, Aug. 2019.
[8] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, “C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject,” J. Phys. Conf. Ser., vol. 1255, no. 012005, pp. 1–7, 2019.
[9] Sudirman, A. P. Windarto, and A. Wanto, “Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia,” IOP Conf. Ser. Mater. Sci. Eng., vol. 420, p. 012089, 2018.
[10] D. Hartama, A. Perdana Windarto, and A. Wanto, “The Application of Data Mining in Determining Patterns of Interest of High School Graduates,” J. Phys. Conf. Ser., vol. 1339, no. 1, 2019.
[11] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” Int. J. Grid Distrib. Comput., vol. 11, no. 8, pp. 87–98, 2018.
[12] S. Harikumar and P. V. Surya, “K-Medoid Clustering for Heterogeneous DataSets,” Procedia Comput. Sci., vol. 70, pp. 226–237, 2015.
[13] A. Buchori, D. Prasetyowati, and Wijayanto, “Design of the magic book math media based on augmented reality,” Test Eng. Manag., 2020.
[14] I. Kamila, U. Khairunnisa, and Mustakim, “Perbandingan Algoritma K-Means dan K-Medoids
untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau,” *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, pp. 119–125, 2019.

[15] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, “Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak,” *CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.

[16] S. Sriyanto, A. Buchori, A. Handayani, P. T. Nguyen, and H. Usman, “Implementation multi factor evaluation process (MFEP) decision support system for choosing the best elementary school teacher,” *Int. J. Control Autom.*, 2020.

[17] S. Defiyanti, M. Jajuli, and N. Rohmawati, “K-Medoid Algorithm in Clustering Student Scholarship Applicants,” *Sci. J. Informatics*, vol. 4, no. 1, pp. 27–33, 2017.

[18] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, “Classification Model C.45 on Determining the Quality of Customer Service in Bank BTN Pematangsiantar Branch,” *J. Phys. Conf. Ser.*, vol. 1255, no. 012002, pp. 1–6, 2019.

[19] S. Sundari, Karmila, M. N. Fadli, D. Hartama, A. P. Windarto, and A. Wanto, “Decision Support System on Selection of Lecturer Research Grant Proposals using Preferences Selection Index,” *J. Phys. Conf. Ser.*, vol. 1255, no. 1, pp. 1–8, 2019.

[20] A. Buchori, P. Setyosari, I. Wayan Dasna, and S. Ulfa, “Developing character building learning model using mobile augmented reality on elementary school student in Central Java,” *Glob. J. Pure Appl. Math.*, 2016.

[21] J. Jamal and D. Yanto, “Analisis RFM dan Algoritma K-Means untuk Clustering Loyalitas Customer,” *Energy*, vol. 9, no. 1, pp. 0–8, 2019.