JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

Classification Based on Configuration Objects by Using Procrustes Analysis

Ridho Ananda[1*], Agi Prasetiadi[2]
[1]Faculty of Industrial Engineering and Design, Institut Teknologi Telkom Purwokerto
[2]Faculty of Informatics, Institut Teknologi Telkom Purwokerto
[1,2]128 D.I. Panjaitan Street, Purwokerto 53147, Indonesia
*Corresponding email: ridho@ittelkom-pwt.ac.id

Received 25 February 2021, Revised 05 May 2021, Accepted 20 May 2021

Abstract — Classification is one of the data mining topics that will predict an object to go into a certain group. The prediction process can be performed by using similarity measures, classification trees, or regression. On the other hand, Procrustes refers to a technique of matching two configurations that have been implemented for outlier detection. Based on the result, Procrustes has a potential to tackle the misclassification problem when the outliers are assumed as the misclassified object. Therefore, the Procrustes classification algorithm (PrCA) and Procrustes nearest neighbor classification algorithm (PNNCA) were proposed in this paper. The results of those algorithms had been compared to the classical classification algorithms, namely k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), AdaBoost (AB), Random Forest (RF), Logistic Regression (LR), and Ridge Regression (RR). The data used were iris, cancer, liver, seeds, and wine dataset. The minimum and maximum accuracy values obtained by the PrCA algorithm were 0.610 and 0.925, while the PNNCA were 0.610 and 0.963. PrCA was generally better than k-NN, SVM, and AB. Meanwhile, PNNCA was generally better than k-NN, SVM, AB, and RF. Based on the results, PrCA and PNNCA certainly deserve to be proposed as a new approach in the classification process.

Keywords – classification, configuration, comparison, data mining, Procrustes

I. INTRODUCTION

Classification is one of the data mining topics quite popular[1]. Classification methods had been implemented in several fields. For example, in the food and agriculture fields, classification was implemented to evaluate food quality [2] and predict soil fertility [3][4]. The method also could be implemented for diagnosing disease [5][6]. for forecasting, it is able to predict weather[7] and the failure of electrical devices[8]. And also, classification had been implemented to assess the performance of employees [9] and students [10].

The principal task of the classification method is the prediction of an object into a certain available group. The prediction process can be performed by using similarity measures, classification trees, or regression. The similarity measures are carried out by k-Nearest Neighbor (k-NN)[11] and Support Vector Machine (SVM)[12]. The classification trees are carried out by AdaBoost (AB)[13] and Random Forest (RF)[14]. Meanwhile, regression approaches are carried out by Logistic Regression (LR)[15] and Ridge Regression (RR)[16].

On the other hand, the Procrustes analysis refers to a technique of matching two configurations. Procrustes analysis formulated computation of the least-squares problem of Y configuration into X configuration by using an orthogonal matrix Q [17]. The first formula of Procrustes was the ordinary Procrustes analysis (OPA). Then, Procrustes had been developed into the Full Procrustes Mean (FPM)[18] and the Goodness-of-fit of Procrustes (GoFP)[19]. Procrustes has recently been implemented in several researchers, namely to determine variables selection[20], measure the quality of biplot analysis[21][22], measure the quality of imputation data[23][24], detect outliers[25], and solve shape clustering problem[26].

Based on the result, Procrustes has a potential to tackle the misclassification problems when the outliers are assumed as the misclassified objects. Therefore, this paper intends to carry out the classification process by using Procrustes. It will become a new strategy where configurations can be utilized as the basis for the classification process. In this paper, there are two Procrustes algorithms proposed in this paper, Procrustes classification (PrCA) and Procrustes nearest neighbor classification algorithms (PNNCA). The concept of k-NN classification is involved in the PNNC. The classification results from the algorithms proposed are compared with the classical classification methods, namely k-NN, SVM, AB, RF, and RR. The data used in this paper are iris, cancer, liver, seeds, and wine dataset.

The difference between this paper and the others is Procrustes' involvement in the classification process at the dataset. It is also hoped that the involvement can contribute to meaningful knowledge, especially in the classification process. This paper is arranged as follows. Section 2 describes a brief history of Procrustes analysis. Furthermore, Section 3 describes the research method used in this paper. Section 4 describes results and discussion. The conclusion is in the last section.

II. A BRIEF HISTORY OF PROCRUSTES ANALYSIS

In ancient Greek, Procrustes' name referred to a bandit who tortured his guests to make a perfect fit with his bed by stretching their limbs or cutting them off. In mathematics, Procrustes referred to a technique of matching two configurations and producing a match measure. Those configurations are matrices of the same size. Suppose **Y** is *n*-by-*p* matrix configuration and **X** is *m*-by-*q* matrix configuration. If $n = m$ and $p < q$ then **Y** needs to be optimally matched to **X** by adding *m*-by-k matrix where $k = q - p$. Similarly, if $n < m$ and $p = q$ then **Y** must be added *l*-by-*p* matrix where $l = m - n$ [22]. To measure the difference between **Y** and **X**, Procrustes utilize the sum of the squared distances E, given by Equation 1.

$$E(\mathbf{Y},\mathbf{X}) = \|\mathbf{Y} - \mathbf{X}\|_F^2 \quad (1)$$

Geometrically, Procrustes works to minimize $E(\mathbf{Y},\mathbf{X})$ by using series of Euclidean similarity transformations, namely translation, rotation, and dilation. Optimal translation in Procrustes is $\mathbf{X}_T = \mathbf{X} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{'}\mathbf{X}$ and $\mathbf{Y}_T = \mathbf{Y} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^{'}\mathbf{Y}$, where n is the number of rows and $\mathbf{1}_n$ is used to denote *n*-by-1 vector having each component equal to 1. Optimal rotation is derived by Ten Berge[27] using the complete form of singular value decomposition (CFSVD) of $\mathbf{X'Y}$, i.e, $\mathbf{X}^{'}\mathbf{Y} = \mathbf{U\Sigma V'}$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_{ij})$ is a real diagonal matrix, **U** and **V** are orthogonal matrices. By using CFSVD, we get solution $\mathbf{Q} = \mathbf{VU'}$, giving the optimal rotation matrix. Optimal dilation is given by scalar $c = \frac{\mathrm{trace}(\mathbf{X'Y})}{\mathrm{trace}(\mathbf{Y'Y})}$.

By using the optimal transformation described above, the ordinary Procrustes analysis (OPA) is given by Equation 2.

$$E_{OPA}(\mathbf{Y},\mathbf{X}) = \|\mathbf{X}_T - c\mathbf{Y}_T\mathbf{Q}\|_F^2 \quad (2)$$

Where $c = \frac{\mathrm{trace}(\mathbf{X'YQ})}{\mathrm{trace}(\mathbf{Y'Y})}$ [18]. The full Procrustes mean (FPM) is a technique for getting the mean of configuration matrices of similar shapes[26]. FPM does not give a measure of the match, so this algorithm abandons in here. Optimal transformation ordering is given by Bakhtiar and Siswadi in the order of translation-rotation-dilation (TRD) as stated in the following theorem[17].

Theorem 1. Given two matrices **X** and **Y** in *n*-by-*p*, the Procrustes between **X** and **Y** after the optimal translation-rotation-dilation (TRD) ordering is given by Equation 3.

$$E_{\mathrm{TRD}}(\mathbf{X},\mathbf{Y}) = \mathrm{trace}(\mathbf{X}_T'\mathbf{X}_T) - \frac{\mathrm{trace}^2(\mathbf{X}_T'\mathbf{Y}_T\mathbf{Q})}{\mathrm{trace}(\mathbf{Y}_T'\mathbf{Y}_T)}. \quad (3)$$

Proof. The complete proof is shown by [17].

Procrustes measure which is given by $E_{OPA}(\mathbf{Y},\mathbf{X})$ or $E_{\mathrm{TRD}}(\mathbf{X},\mathbf{Y})$ does not comply with the symmetrical property where $E(\mathbf{X},\mathbf{Y}) \neq E(\mathbf{Y},\mathbf{X})$. For the problem, Bakhtiar and Siswadi in [19] has embedded the symmetrical property by adding other transformation namely a normalization as stated in the following theorem

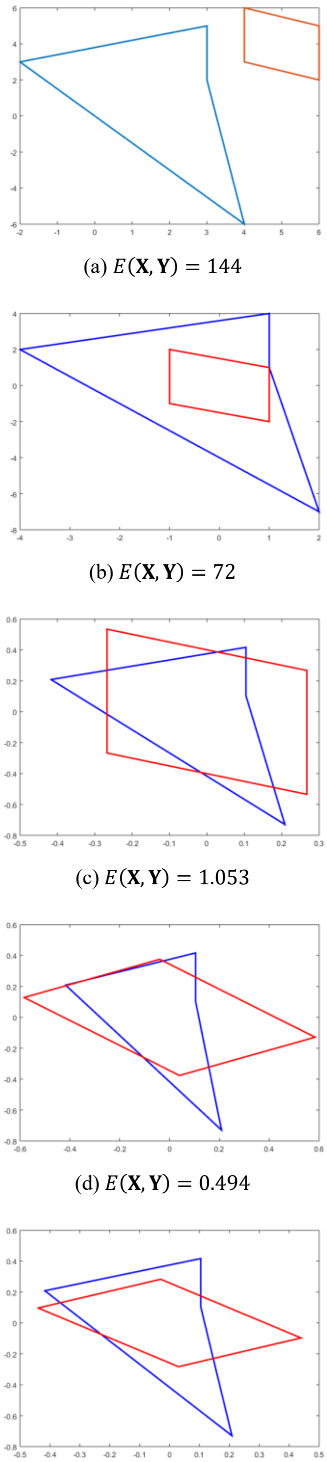
Theorem 2. Given two matrices **X** and **Y** in *n*-by-*p*, the Procrustes between **X** and **Y** after the optimal translation-normalization-rotation-dilation ordering comply with the symmetrical property given by Equation 4.

$$E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y}) = E_{\mathrm{TNRD}}(\mathbf{Y},\mathbf{X}) = 1 - \left(\sum_{i=1}^{r} \sigma_{ii}\right)^2, \quad (4)$$

where r and σ_{ii} are rank and singular value of $\bar{\mathbf{X}}_T'\bar{\mathbf{Y}}_T$ or $\bar{\mathbf{Y}}_T'\bar{\mathbf{X}}_T$ with $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$ are matrices after normalization process by using formula $\bar{\mathbf{X}} = \frac{\mathbf{X}}{\|\mathbf{X}\|_F}$ and $\bar{\mathbf{Y}} = \frac{\mathbf{Y}}{\|\mathbf{Y}\|_F}$.

Proof. The complete proof is shown by [19].

Another fact of $E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y})$ is the value of $E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y})$ among 0 and 1, $0 \leq E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y}) \leq 1$. If $E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y}) \approx 0$ then it means that **X** and **Y** have an excellent match. Conversely, if $E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y}) \approx 1$ then **X** and **Y** have the poor match. Based on the fact, Bakhtiar and Siswadi has made the opposite interpretation that is $\mathrm{GoFP}(\mathbf{X},\mathbf{Y}) = 1 - E_{\mathrm{TNRD}}(\mathbf{X},\mathbf{Y}) = (\sum_{i=1}^{r} \sigma_{ii})^2$ where it could be judged as a goodness-of-fit of the best matching or goodness-of-fit of Procrustes. The illustration of Procrustes is shown in Fig.1.



(a) $E(\mathbf{X},\mathbf{Y}) = 144$

(b) $E(\mathbf{X},\mathbf{Y}) = 72$

(c) $E(\mathbf{X},\mathbf{Y}) = 1.053$

(d) $E(\mathbf{X},\mathbf{Y}) = 0.494$

(e) $E(\mathbf{X},\mathbf{Y}) = 0.433$

Fig. 1. Those Figures Show Procrustes Measure (a) Before Transformation Process, (b) Translation, (c) Translation-Normalization, (d) Translation-Normalization-Rotation, and (e) Translation-Normalization-Rotation-Dilation.

III. RESEARCH METHODS

A. *Procrustes Classification Algorithm*

The basic idea of the Procrustes classification algorithm (PrCA) is a change in the group's configuration because of the testing data entry. If the testing data is entered into a particular available group, it will change its configuration. If the change is the largest, it can be assumed that the testing data is misclassified in the group. The visualization of this concept is given in Fig.2.

From Fig.2, we are intuitively convinced that the configurations of **X** and **Y** are different. Dissimilarity measures of those configurations can be obtained using the GoFP. If GoFP is close to 1, then the difference between **X** and **Y** is tiny. It means that the entry of testing data does not change the initial configuration significantly, so it can be assumed that the testing data is part of the group. Conversely, if GoFP is close to 0, then the difference between **X** and **Y** is huge. It means that the entry of testing data changes the initial configuration significantly, so it can be assumed that the testing data is not part of the group.

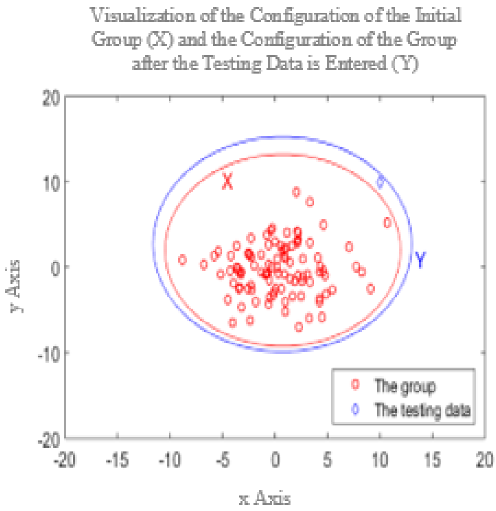


Fig. 2. Visualization of the Configuration of the Initial Group (**X**) and the Configuration of the Group after the Testing Data is Entered (**Y**)

The next problem arises when the GoFP will be calculated. Suppose that **X** is *n*-by-*p* matrix, so **Y** is certainly *(n+1)*-by-*p* matrix. There is a difference in the size of **X** and **Y**. As a result, GoFP can not calculate. To solve the problem, we have to add one object in **X** where it does not change the configuration of **X** extremely. One of the solutions is to select a particular object from **X**. In this paper, it will be selected from the prototype of **X** . Based on that experience, the Procrustes classification algorithm is proposed with the following steps.

1. Suppose that $\mathbf{X}_i$ *n*-by-*p* is a matrix of the *i*th group ($i = 1,2,\dots,k$), and **a** is testing data.
2. we build $\mathbf{Y}_i = \begin{bmatrix}\mathbf{X}_i \\ \mathbf{a}'\end{bmatrix}$ with size *(n+1)*-by-*p*.

3. Add $\mathbf{X}_i$ with its prototype to be optimally matched to $\mathbf{Y}_i$.
4. Compute $\mathrm{GoFP}(\mathbf{X}_i, \mathbf{Y}_i)$, **a** is classified into $\mathbf{X}_i$ if
$$\mathrm{GoFP}(\mathbf{X}_i, \mathbf{Y}_i) \geq \mathrm{GoFP}(\mathbf{X}_j, \mathbf{Y}_j)$$
For every $j \in \{1,2,\dots,k\}$.
5. Perform the above procedure for each testing data.

B. Procrustes Nearest Neighbor Classification Algorithm

The basic idea of the Procrustes nearest neighbor classification algorithm (PNNCA) is that add $\mathbf{X}_i$ with object from $\mathbf{X}_i$ closest to the testing data. To be optimally matched to $\mathbf{Y}_i$, we have to add $\mathbf{X}_i$ with its prototype on PrCA. Now, we try to add $\mathbf{X}_i$ with another object from $\mathbf{X}_i$. K-NN algorithm works by seeing the nearest neighbor. Each of the objects in all groups is a neighbor of testing data. Based on all neighbors, testing data will be classified into the group which contains its k nearest neighbor. Based on this fact, we know that each object in a certain group is a neighbor of testing data. One of them has the smallest distance with testing data, suppose **b** where $\mathbf{b} \in \mathbf{X}_i$. If $\mathbf{X}_i$ is added **b** then we have added $\mathbf{X}_i$ so it has the same size with $\mathbf{Y}_i$ using the concept of k-NN.

C. Research Flow

The research flow used in this paper consists of four main steps: preprocessing of data, classification process, computation of accuracy of the classification results, and comparison of the classification results. Data used are iris, cancer, liver, seeds, and wine datasets obtained from the UCI website. In preprocessing of data, data will be standardized by using zscore because there are features that have different units. In the classification process, classification algorithms used are PrCA, PNNCA, k-NN, SVM, AB, RF, LR, and RR. The testing data are obtained by using k-fold cross-validation ($k = 10$). K-fold cross-validation will divide data into ten parts randomly[28]. Each part will be testing data consecutively, and the remaining part will be training data. The accuracy results from an algorithm are obtained from the average accuracy of all testing data by using (5).

$$\text{Accuracy} = \frac{\text{A number of true classification}}{\text{A number of processed data}}, \quad (5)$$

The classification results of each algorithm are compared to get the best result. In simple terms, the research flow is shown in Fig.3.

IV. RESULTS

A. The data used

The data used in this research are secondary data obtained from the website UCI, namely iris, cancer, liver, seeds, and wine. All of these data are quantitative data. The description of these data used is shown in Table 1. The table gives information about the number of objects, features, and classes in each data.

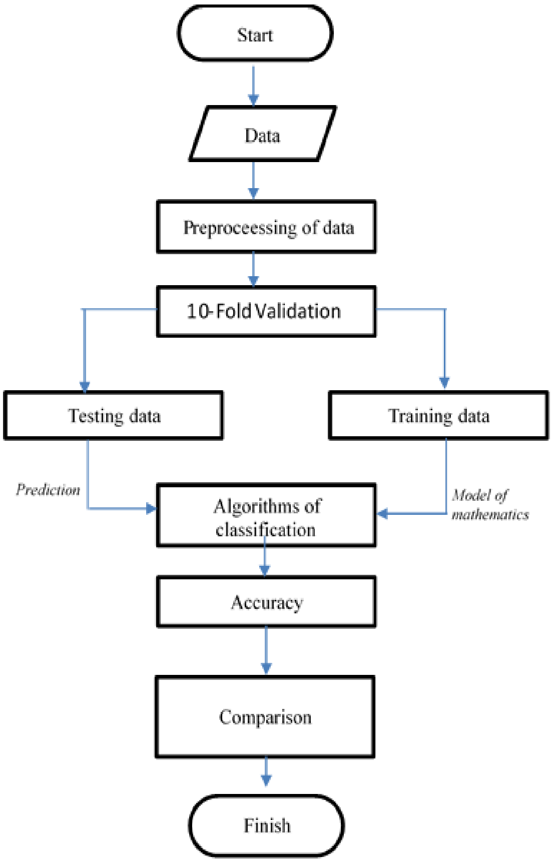


Fig. 3. The Research Flow In This Research

Table 1. Description of the Dataset

No	Dataset	*n* objects	*n* Features	*n* class
1	Iris	150	5	3
2	Cancer	116	10	2
3	Liver	345	7	2
4	Seeds	210	8	3
5	Wine	178	14	3

Table 2 shows the maximum and minimum variance (S^2) of dataset. From the table, we know that cancer, liver, seeds, and wine dataset have different S^2 min and S^2 max significantly. It shows that those datasets have features with different units. These differences will certainly affect the classification results, where features with a large variance will be more influential. To overcome it, those data are standardized first by using z-score . Whereas iris dataset have not any different S^2 min and S^2 max significantly. It shows that the dataset has features with the same units, so it does not need to be standardized.

Table 2. Standard Deviation Each Dataset Used

No	Dataset	*S* min	*S* maks
1	Iris	0.19	3.11
2	Cancer	13.26	119,655.53
3	Liver	11.14	1,540.92
4	Seeds	0.001	8.47
5	Wine	0.01	99,166.73

B. The Classification Results

The getting process of the accuracy results is obtained from each algorithm iteration 100 times in each dataset. It is done to see the convergence of the accuracy results by using (6).

$$a = \lim_{n\to\infty} a_n, \qquad (6)$$

where a_n is accuracy result in nth iteration, and a is the convergence of the accuracy[29]. Computation in this paper uses Matlab.

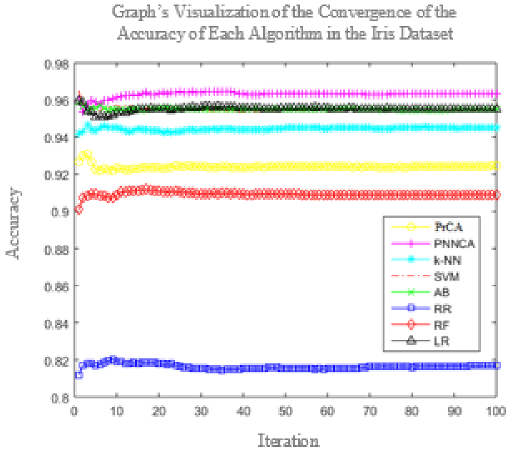


Fig.4. Graph's Visualization of the Convergence of the Accuracy of Each Algorithm in the Iris Dataset

Figure 4 shows the accuracy convergence from PrCA, PNNCA, k-NN, SVM, AB, RR, RF, and LR algorithms in the iris dataset. From the chart, we know that each algorithm's accuracy results are satisfactory because the minimal value of the accuracy is above 0.800. The average accuracy of each algorithm precisely shows in Table 3. In Table 3, we know that the best algorithms are PNNCA and all algorithms' accuracy results are quite similar.

Table 3. The Average Accuracy of Each Algorithm in the Iris Dataset

Algorithm	Iris Dataset
PrCA	0.9245
PNNCA	0.9633
k-NN	0.9450
SVM	0.9545
AB	0.9551
RR	0.8170
RF	0.9090
LR	0.9550

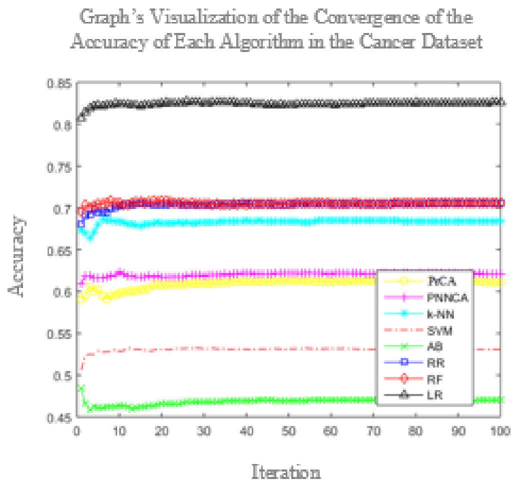


Fig.5. Graph's Visualization of the Convergence of the Accuracy of Each Algorithm in the Cancer Dataset

Figure 5 shows the chart of the convergence of the accuracy results of each algorithm in the cancer dataset. We know intuitively that the AB algorithm result is below 0.500. Its result is certainly not satisfactory because the true classification is less than the wrong classification. If we see the other algorithms, it knows that all algorithms, except AB, have good enough because of the dominant true classification results. The average accuracy of each algorithm precisely shows in Table 4.

Table 4. The Average Accuracy of Each Algorithm in the Cancer Dataset

Algorithm	Cancer Dataset
PrCA	0.6105
PNNCA	0.6207
k-NN	0.6844
SVM	0.5307
AB	0.4703
RR	0.7059
RF	0.7058
LR	0.8256

Table 4 shows exactly the accuracy values of each algorithm. From the table, we know that the result of PrCA and PNNCA is quite similar; the difference is only 0.0102. we also know that PrCA and PNNCA results are good enough because the true classification is more than the false classification. However, those algorithms are not the best algorithm in the cancer dataset.

Figure 6 shows the chart of the convergence of the accuracy from all algorithms used in the liver dataset. We find again that there is one algorithm whose accuracy results are below 0.500. The algorithm is SVM. However, other algorithms are good enough because of the dominant true classification results. To know the results of each algorithm exactly needs to see Table 5.

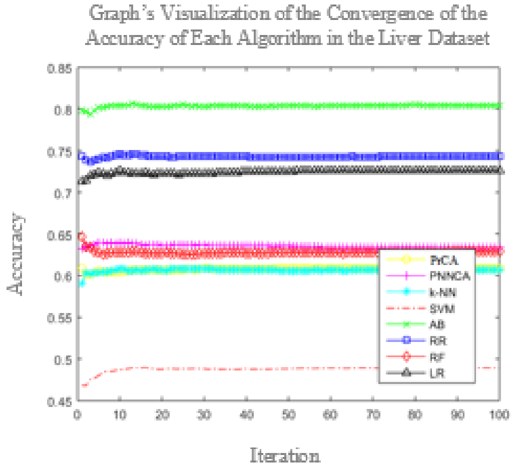


Fig.6. Graph's Visualization of the Convergence of the Accuracy of Each Algorithm in the Liver Dataset

Table 5. The Average Accuracy of Each Algorithm in the Liver Dataset

Algorithm	Liver dataset
PrCA	0.6093
PNNCA	0.6334
k-NN	0.6069
SVM	0.4890
AB	0.8039
RR	0.7434
RF	0.6285
LR	0.7261

Table 5 shows the average value of the accuracy of each algorithm in the liver dataset with 100 times repetition. From the table, we know that PrCA and PNNCA results are better than k-NN and SVM. It is clearly that PNNCA is better than PrCA in the liver dataset. Although PrCA and PNNCA results are not the best, their results are good enough because the dominant classification is true.

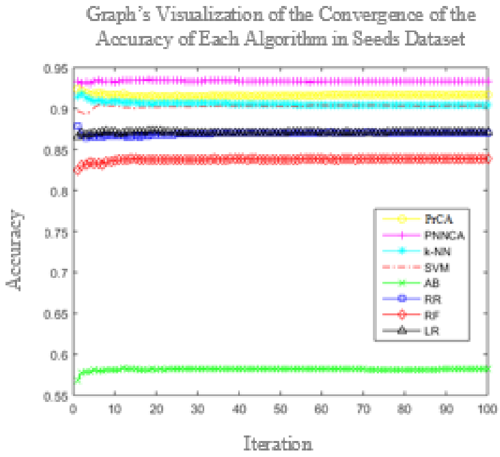


Fig.7. Graph's Visualization of the Convergence of the Accuracy of Each Algorithm in Seeds Dataset

The convergence of the accuracy results in the seeds dataset shown in Fig.7 shows that all algorithms, except Adaboost (AB), are satisfactory because their accuracy is above 0.800. In comparison, the AB results are only good enough because its accuracy is slightly above 0.550. The precise results of each algorithm can be known in Table 6.

Table 6. The Average Accuracy of Each Algorithm in the Seeds Dataset

Algorithm	Seeds Dataset
PrCA	0.9163
PNNCA	0.9331
k-NN	0.9040
SVM	0.9034
AB	0.5816
RR	0.8706
RF	0.8387
LR	0.8708

From Table 6, we know that the PrCA and PNNCA results are satisfactory because the accuracy value are above 0.900. Moreover, we also know that PNNCA is the best algorithm, while PrCA is the second-best algorithm in the seeds dataset.

Figure 8 shows intuitively that the SVM results are about 0.500, and the k-NN results are about 0.600 in the wine dataset. While other algorithms have accuracy results is about or above 0.800. To know exactly needs to see Table 7.

From Table 7, we know the average accuracy of each algorithm used. The SVM algorithm result is 0.5004, so the result is good enough. If we focus on PrCA PNNCA, we will see that their results are satisfactory in the wine dataset because those are above 0.800, although one of them is not the best algorithm. We also know that the PNNCA is better than PrCA.

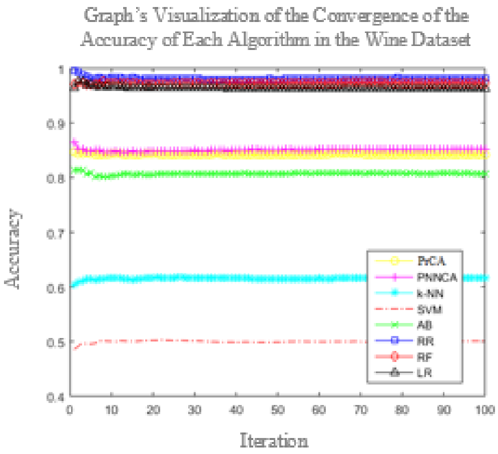


Fig.8. Graph's Visualization of the Convergence of the Accuracy of Each Algorithm in the Wine Dataset

Table 7. The Average Accuracy of Each Algorithm in the Wine Dataset

Algorithm	Wine Dataset
PrCA	0.8421
PNNCA	0.8528
k-NN	0.6166
SVM	0.5004
AB	0.8075
RR	0.9819
RF	0.9733
LR	0.9619

V. DISCUSSION

The results that have been shown previously provide some necessary information about the classification algorithm comparisons. First, PrCA is better than k-NN in the liver, seeds, and wine dataset, while PNNCA is better than k-NN in the iris, liver, seeds, and wine dataset. Second, PrCA is better than SVM in cancer, liver, seeds, and wine dataset, while PNNCA is better than SVM in all datasets used. Third, PrCA is better than AB in the cancer, seeds, and wine dataset, while PNNCA is better than AB in iris, cancer, seeds, and wine. Fourth, PrCA and PNNCA are better than RR in the same dataset, namely iris and seeds. Fifth, PrCA is better than RF in the iris and seeds dataset, while PNNCA is better than RF in iris, liver, and seeds.

At last, PrCA is just better than LR in the seeds dataset, while PNNCA is better than LR in iris and seeds. The facts show that PrCA has good results predominantly compared to k-NN, SVM, and AB. While PNNCA has good results predominantly compared to k-NN, SVM, AB, and RF. We also know that selected object to add $\mathbf{X}_i$ for optimally matching to $\mathbf{Y}_i$ give impact in the classification results, where PNNCA is better than PrCA in this case. In general, the results from PNNCA and PrCA show that the involvement of all objects in the group in the classification process based on similarity measure is more advantageous than only several objects from the group.

VI. CONCLUSION

This paper discussed the classification algorithms proposed by using Procrustes, namely PrCA and PNNCA. The results conclude that the results of PrCA are quite similar to PNNCA, but PNNCA is better than PrCA in all datasets used. PrCA has outperformed three of the six comparing algorithms, which are k-NN, SVM, and AB. Meanwhile, PNNCA has outperformed four of them that are k-NN, SVM, AB, and RF. Based on the results, PrCA and PNNCA certainly deserve to be proposed as a new approach in the classification process.

ACKNOWLEDGEMENTS

We thank the support of Ristekdikti through a 2020 external grant for the financial support of the research and IT Telkom Purwokerto for the support for this paper.

REFERENCES

[1] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for kNN Classification,” *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, 2017, doi: 10.1145/2990508.

[2] A. M. Jiménez-Carvelo, A. González-Casado, M. G. Bagur-González, and L. Cuadros-Rodríguez, “Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review,” *Food Res. Int.*, vol. 122, no. February, pp. 25–39, 2019, doi: 10.1016/j.foodres.2019.03.063.

[3] V. Rajeswari and K. Arunesh, “Analysing soil data using data mining classification techniques,” *Indian J. Sci. Technol.*, vol. 9, no. 19, 2016, doi: 10.17485/ijst/2016/v9i19/93873.

[4] E. M. S. Djodiltachoumy, “Analysis of Data Mining Techniques for Agriculture Data,” *Indian J. Sci. Technol.*, vol. 4, no. 2, pp. 1311–1313, 2016, doi: 10.17485/ijst/2016/v9i38/101962.

[5] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[6] S. Anwar Lashari, R. Ibrahim, N. Senan, and N. S. A. M. Taujuddin, “Application of Data Mining Techniques for Medical Data Classification: A Review,” *MATEC Web Conf.*, vol. 150, pp. 1–6, 2018, doi: 10.1051/matecconf/201815006003.

[7] R. Prasetya and A. Ridwan, “Data Mining Application on Weather Prediction Using Classification Tree, Naïve Bayes and K-Nearest Neighbor Algorithm With Model Testing of Supervised Learning Probabilistic Brier Score, Confusion Matrix and ROC,” *J. Appl. Commun. Inf. Technol.*, vol. 4, no. 2, pp. 25–33, 2019.

[8] A. Contreras-Valdes, J. P. Amezquita-Sanchez, D. Granados-Lieberman, and M. Valtierra-Rodriguez, “Predictive data mining techniques for fault diagnosis of electric equipment: A review,” *Appl. Sci.*, vol. 10, no. 3, pp. 1–24, 2020, doi: 10.3390/app10030950.

[9] J. M. and C. A., “Application of Data Mining Classification in Employee Performance Prediction,” *Int. J. Comput. Appl.*, vol. 146, no. 7, pp. 28–35, 2016, doi: 10.5120/ijca2016910883.

[10] A. Ashraf, S. Anwer, and M. G. Khan, “A Comparative Study of Predicting Student ’ s Performance by use of Data A Comparative Study of Predicting Student ’ s Performance by use of Data Mining Techniques,” *Am. Sci. Res. J. Eng. Technol. Sci.*, vol. 44 No.1, no. October, pp. 122–136, 2018.

[11] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.

[12] C. C. V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995, doi: 10.1109/64.163674.

[13] N. K. K. N. S. P. K. Y. V. N. H. Deekshitulu, “Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System,” *Int. J. Inf. Sci. Tech.*, vol. 2, no. 3, pp. 63–75, 2012, doi: 10.1523/JNEUROSCI.4623-04.2005.

[14] T. K. Ho, “Random decision forests,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi: 10.1109/ICDAR.1995.598994.

[15] J. Cramer, “the origin of logistic regression,” 2002. [Online]. Available: https://papers.tinbergen.nl/02119.pdf.

[16] K. Rakesh and P. N. Suganthan, “An Ensemble of Kernel Ridge Regression for Multi-class Classification,” *Procedia Comput. Sci.*, vol. 108, pp. 375–383, 2017, doi: 10.1016/j.procs.2017.05.109.

[17] T. Bakhtiar and Siswadi, “Orthogonal procrustes analysis: Its transformation arrangement and minimal distance,” *Int. J. Appl. Math. Stat.*, vol. 20, no. M11, pp. 16–24, 2011.

[18] I. L. D. K. V. Mardia, *Statistical Shape Analysis with applications in R*, 2nd ed. 2016.

[19] T. S. Bakhtiar, “ON THE SYMMETRICAL PROPERTY OF PROCRUSTES MEASURE OF DISTANCE,” vol. 99, no. 3, pp. 315–324, 2015.

[20] A. Muslim and T. Bakhtiar, “Variable selection using principal component and procrustes analyses and its application in educational data,” *J. Asian Sci. Res.*, vol. 2, no. 12, pp. 856–865, 2012, [Online]. Available: http://www.aessweb.com/pdf-files/856-865.pdf.

[21] Siswadi and T. Bakhtiar, “Goodness-of-fit of biplots via procrustes analysis,” *Far East J. Math. Sci.*, vol. 52, no. 2, pp. 191–201, 2011.

[22] Siswadi, T. Bakhtiar, and R. Maharsi, “Procrustes analysis and the goodness-of-fit of biplots: Some thoughts and findings,” *Appl. Math. Sci.*, vol. 6, no. 69–72, pp. 3579–3590, 2012.

[23] R. Ananda, Siswadi, and T. Bakhtiar, “Goodness-of-Fit of the Imputation Data in Biplot Analysis,” *Far East J. Math. Sci.*, vol. 103, no. 11, pp. 1839–1849, 2018, doi: 10.17654/ms103111839.

[24] R. Ananda, A. R. Dewi, and N. Nurlaili, “a Comparison of Clustering By Imputation and Special Clustering Algorithms on the Real Incomplete Data,” *J. Ilmu Komput. dan Inf.*, vol. 13, no. 2, pp. 65–75, 2020, doi: 10.21609/jiki.v13i2.818.

[25] F. Novika and T. Bakhtiar, “The Use of Biplot Analysis and Euclidean Distance with Procrustes Measure for Outliers Detection,” *Int. J. Eng. Manag. Res. Page Number*, no. 1, pp. 194–200, 2018, [Online]. Available: www.ijemr.net.

[26] K. Iwata, *Shape clustering as a type of procrustes analysis*, vol. 11304 LNCS. Springer International Publishing, 2018.

[27] J. M. F. Ten Berge, “The rigid orthogonal procrustes rotation problem,” *Psychometrika*, vol. 71, no. 1, pp. 201–205, 2006, doi: 10.1007/s11336-004-1160-5.

[28] H. Azis, P. Purnawansyah, F. Fattah, and I. P. Putri, “Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung,” *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 81–86, 2020, doi: 10.33096/ilkom.v12i2.507.81-86.

[29] M. Taboga, *Lectures on Probability Theory and Mathematical Statistics*, 3rd ed. CreateSpace Independent Publishing Platform, 2017.