Hierarchical Clustering Algorithm Based on Density Peaks using Kernel Function for Thalassemia Classification

S Hartini¹ and Z Rustam¹
¹Department of Mathematics, University of Indonesia, Depok 16424, Indonesia,

Email: sri.hartini@sci.ui.ac.id, rustam@ui.ac.id

Abstract. Thalassemia is an inherited blood disorder and one of the most common genetic diseases in the world, with its classification due to the processes involved in patients’ treatment. Therefore, it is essential to obtain a precise diagnosis due to an appropriate treatment receives a higher lifetime value. A new method based on the kernel, modified from hierarchical clustering based on density peaks (HCDP) was proposed in this research. Furthermore, the concept of $k$-nearest neighbor and hierarchical clustering, consists of three steps were utilized. These steps include local density calculation, hierarchy representation, and optimal cluster extraction. The modification of this technique is therefore based on the polynomial kernel function, which was expected to be more accurate in separating data that cannot be detached linearly. Data utilized was obtained from Harapan Kita Hospital, West Jakarta, Indonesia, and it consists of 82 thalassemia and 68 non-thalassemia samples. Using the proposed method, the performance of HCDP with/without kernel function in this paper were examined using 10-fold cross-validation and compared using the confusion matrix by calculating its F1-Score. The results concluded that hierarchical clustering based on density peaks gives approximately 67.77 percent F1-Score, while 70.06 percent is obtained when the method is combined with the kernel function.

1. Introduction

Thalassemia, which is derived from two Greek words *thalassa* (sea) and *haima* (blood), is the most common genetic disorders that are inherited [1]. This disease is prevalent in most countries of the world such as the Mediterranean, Central Asia, Middle East, Southern China, India, Far East, and also the countries along the northern coast of South America and Africa [2]. The total incidence of patients with this disease each year is estimated at 1:100,000 in every part of the world, and 1:10,000 people in the European Union [3].

According to the Thalassemia International Federation, it is estimated that 7% of the population in the world is a carrier of a severe hemoglobin disorder including thalassemia and 300-500,000 children are born with it each year. Furthermore, there were also reports that only 200,000 people living with this disease are registered and receiving regular treatment around the world [4]. Thalassemia classification, therefore, aims to give advance research in clinical care for precise diagnosis. This classification enables the delivery of appropriate treatment capable of improving patients’ quality of life and increasing their survival rate.
Numerous classification studies have previously been carried out, involving various machine learning techniques. Paokanta et al. [5] utilized some methods such as $k$-Nearest Neighbors (KNN), Bayesian Networks (BNs), Multi-Layer Perceptron (MLP), Multinomial Logistic Regression, and also Naïve Bayes with the accuracy of 80 percent. Some neural networks and decision trees also used red blood cells, reticulocytes, and platelets characteristics [6]. Using the Complete Blood Count (CBC) test and artificial intelligence algorithms, Yousefian et al. [7] found that red blood cells, hemoglobin, mean corpuscular volume, and hematocrit are essential characteristics in classifying thalassemia.

Hierarchical clustering based on density peaks has proven more accurate for several complex datasets [8]. Therefore, in this study, hierarchical clustering based on density peaks using kernel function was adopted and expected to be more accurate in the classification.

2. Material and Methods

2.1. Material

Thalassemia dataset was obtained from Harapan Kita Hospital in West Jakarta, Indonesia. This dataset consists of 82 thalassemia and 68 non-thalassemia samples with numerical characteristics described by ten attributes, as shown in Table 1.

| Attribute Name | Description |
|----------------|-------------|
| HGB            | The number of hemoglobin in grams per deciliter of blood |
| HCT            | Hematocrit or the volume percentage of red blood cells |
| WBC            | The number of white blood cells per microliter of blood |
| Basophils      | The basophils percentage in the blood |
| Eosinophils    | The eosinophils percentage in the blood |
| Rod neutrophils| The bands (almost mature) neutrophils percentage with rod-shaped |
| Segmented neutrophils | The segmented (fully mature) neutrophils percentage in the blood |
| Lymphocytes    | The lymphocytes percentage in the blood |
| Monocytes      | The monocytes percentage in the blood |
| PLT            | The number of platelet count per microliter of blood |

2.2. Methods

This study uses a hierarchical clustering method based on density peaks with/without kernel function. The performances of these two methods are evaluated using 10-fold cross-validation and later compared.

2.2.1. Kernel Function. The kernel function for every $x, y \in R^n$ is defined as below [9]

$$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

(1)

where $\phi$ is the function for mapping the input data to the feature space. Meanwhile, the distance between two points is defined as follows [9]

$$\text{dist}(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}$$

(2)

This kernel function representation maps the data into a high dimensional feature space useful in solving complex real-world applications which cannot be solved by linear functions [10]. The kernel function can be linear, polynomial, and Gaussian radian basis kernel function [11]. However, polynomial kernel function with the formula expressed in equation (3) is used in this research. The polynomial degree used can be any value, but for simplicity, only the first ten polynomial with degree $p$ is examined.

$$K(x, y) = (x \cdot y + 1)^p$$

(3)
2.2.2. **Hierarchical Clustering Based on Density Peaks.** There are three steps in this method: \(k\)-nearest density calculations, hierarchy illustration in a graph, and cluster optimization, which is used to obtain the final clusters. \(k\)-nearest density calculation uses \(k\)-nearest neighbors whose \(i\)’s density defined as below.

\[
\rho_i = (i,j) \tag{4}
\]

where \(kNN(i) = \{j \in D | d(i,j) \leq dist(i,NN_k(i))\}\) and \(NN_k(i)\) is the \(k\)th nearest point to \(i\) by \(dist(.)\) computed using standard Euclidean distance [8]. In order to implement \(k\)-nearest neighbors, KD tree was used to compute the distance efficiently. It is defined as the multidimensional binary search trees in which data structure is divided into \(k\) partition search space [12].

After computing \(k\)-nearest density \(\rho_i\), its distance \(\delta_i\) from points with a higher density is calculated as below [8].

\[
\delta_i = dist(i,j) \tag{5}
\]

The hierarchical clustering was used based on the density peaks to obtain a tree which gives distance and density information. It was conducted by treating data as vertices in a graph. However, each vertex not related to the maximal density was connected with an edge weight \(\delta_i\) to its nearest neighbor of higher density [8]. The edge was then sorted and iteratively deleted in decreasing order of weight. After that, a clustering tree containing all partitions was obtained.

Finally, the problem was handled as an optimization issue to maximize the total stabilities of the clusters shown below [8].

\[
S = \sum_{i=1}^{K} S(C_i) \tag{6}
\]

where \(K\) is defined as the number of clusters and \(S(C_i)\) as [8].

\[
S(C_i) = \sum_{x_j \in C_i} \left( \frac{1}{\delta_{exclude}(x_j,C_i)} - \frac{1}{\delta_{emerge}(C_i)} \right) \tag{7}
\]

where \(\delta_{exclude}\) is the maximum weight with deleting edges when excluding point \(x_j\) from the cluster and \(\delta_{emerge}\) being the maximum weight whose cluster \(C_i\) was separated from the previous one [8].

2.2.3. **Hierarchical Clustering Based on Density Peaks using Kernel Function.** When the kernel function was used, there was not much difference with the original. There were also three steps in hierarchical clustering based on density peaks using kernel function. These include \(k\)-nearest density calculations, hierarchy representation in a graph, and cluster optimization. The difference exists due to the substitution of the distance formula with kernel function, thereby, making it possible to compare the performance of these two methods in the classification. \(k\)-nearest density calculation uses \(k\)-nearest neighbors \(i\)’s density is defined as

\[
\rho_i = \sqrt{K(i,i) - 2K(i,j) + K(j,j)} \tag{8}
\]

where \(kNN(i) = \{j \in D | K(i,j) \leq K(i,NN_k(i))\}\) and \(NN_k(i)\) is the \(k\)th nearest point to \(i\) by \(dist(.)\). After computing its \(k\)-nearest density \(\rho_i\), its distance \(\delta_i\) from points with a higher density is calculated as follows:

\[
\delta_i = \sqrt{K(i,i) - 2K(i,j) + K(j,j)} \tag{9}
\]
The hierarchical clustering based on density peaks using kernel function was used to obtain a tree conducted by treating data as vertices in a graph. Each vertex besides the one for maximum density was connected with an edge weight $\delta_i$ to its nearest neighbor of higher density. The edge was then sorted and iteratively removed in decreasing order of weight, and the clustering problem was transformed into a sub-tree partition problem. Finally, the optimal cluster extraction was carried out in the same way as the hierarchical clustering based on density peaks without using kernel function.

2.2.4. Performance Measure. F1-score was used in this paper to evaluate the methods based on the confusion matrix. It has a balance interest in precision and sensitivity, which distinguishes the correct classification of labels within different classes [13]. They are the measures used to measure how well the method obtained relevant information [14]. The formula of F1-Score is shown below.

$$F1\text{-Score} = \frac{2 \times \text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}} \times 100\% \quad (10)$$

with the sensitivity and precision defined below.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \quad (11)$$

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\% \quad (12)$$

in which TP is the number of thalassemia samples correctly diagnosed, FN is the number of thalassemia samples incorrectly diagnosed, and FP is the number of non-thalassemia samples incorrectly diagnosed [15].

3. Results and Discussions
The performance of hierarchical clustering based on density peaks with/without kernel function was examined using 10-fold cross-validation. In this, the dataset was split randomly into ten mutually exclusive folds of approximately equal size [16]. Furthermore, each fold was taken as the validation data for testing the classification model, and this process was repeated ten times, which was advantageous to the repeated random sub-sampling [17]. The performance of hierarchical clustering based on density peaks is given in Table 2 below.

| Table 2. Performance measurements of hierarchical clustering based on density peaks |
|---|---|---|
| Fold | F1-Score (%) | Running Time (s) |
| 1 | 69.43 | 5.59 |
| 2 | 67.65 | 1.11 |
| 3 | 69.57 | 1.26 |
| 4 | 68.06 | 0.98 |
| 5 | 68.75 | 0.83 |
| 6 | 66.32 | 1.00 |
| 7 | 66.67 | 1.18 |
| 8 | 69.43 | 0.95 |
| 9 | 63.44 | 1.01 |
| 10 | 68.42 | 1.36 |

The average value is calculated, with the F1-Score and running time of the hierarchical clustering based on density peaks method acquired is 67.77 percent and 1.53 seconds, respectively. The first ten polynomial degrees are examined with the kernel using the same 10-fold cross-validation. The
performance of hierarchical clustering based on density peaks using the polynomial kernel is given in Table 3.

**Table 3. Performance measurements of hierarchical clustering based on density peaks using a polynomial kernel function**

| Polynomial Degree | Average of F1-Score (%) | Average of Running Time (s) |
|-------------------|-------------------------|----------------------------|
| 1                 | 67.77                   | 1.26                       |
| 2                 | 55.32                   | 1.17                       |
| 3                 | 68.29                   | 1.17                       |
| 4                 | 68.37                   | 1.25                       |
| 5                 | 69.55                   | 1.25                       |
| 6                 | 69.55                   | 1.27                       |
| 7                 | 69.55                   | 1.20                       |
| 8                 | 70.06                   | 1.29                       |
| 9                 | 70.06                   | 1.28                       |
| 10                | 70.06                   | 1.16                       |

According to Table 3, we can see that the average of F1-Score and also the running time varied according to the polynomial degree used, as illustrated in Figure 1 and Figure 2 below.

![Figure 1](image)

**Figure 1.** The average of F1-Score of hierarchical clustering based on density peaks using polynomial kernel functions
Figure 2. The average of the running time of hierarchical clustering based on density peaks using polynomial kernel functions.

The highest average value of F1-Score is 70.06 percent, which was obtained using the eighth, ninth, tenth-degree polynomial kernel. Furthermore, these F1-Scores were obtained after 1.29 seconds, 1.28 seconds, and 1.16 seconds. Therefore, tenth-degree polynomial kernel gives better performance in average running time. The performance of hierarchical clustering based on density peaks using tenth-degree polynomial is given in Table 4.

Table 4. Performance measurements of hierarchical clustering based on density peaks using a tenth-degree polynomial kernel function

| Fold | F1-Score (%) | Running Time (s) |
|------|--------------|------------------|
| 1    | 71.43        | 1.14             |
| 2    | 68.29        | 1.16             |
| 3    | 70.19        | 1.16             |
| 4    | 69.57        | 1.14             |
| 5    | 70.19        | 1.16             |
| 6    | 68.29        | 1.24             |
| 7    | 69.57        | 1.14             |
| 8    | 72.04        | 1.16             |
| 9    | 70.19        | 1.17             |
| 10   | 70.81        | 1.15             |

In this table, the F1-Score reaches up to 70 percent in several folds using this polynomial kernel function, with 72.04 percent as the highest F1-Score obtained.

4. Conclusion
Thalassemia, as an inherited blood disorder, is the most common genetic disorders in the world. It is necessary to classify it since the initial process of executing appropriate treatment is capable of improving patients’ quality of life and increase their survival rate. The dataset was obtained from Harapan Kita Hospital, West Jakarta, Indonesia. This dataset consists of 82 thalassemia and 68 non-
thalassemia samples which were used to evaluate the performance of hierarchical clustering based on density peaks with/without the kernel function.

These two methods were examined using 10-fold cross-validation. From the experiments, hierarchical clustering based on density peaks gives approximately 67.77 percent F1-Score in 1.53 seconds. However, 70.06 percent of F1-Score is obtained when combined with a polynomial kernel function. These values are relatively good, but still, need some improvement, although both methods had good running time, which is less than 2 seconds for almost all experiments.

It is required that further research be carried out using the new model in order to improve their performance. Other kernel parameters can also be used for the same purpose, with the possibility to observe other performance measurements. Higher F1-Score is then expected to give more accurate and precise classification to obtain a better diagnosis.

Acknowledgments
This research was financially supported with a PDUPT 2019 research grant scheme (ID number NKB-1621/UN2.R3.1/HKP.05.00/2019) by the Indonesian Ministry of Research and Higher Education. The authors are grateful to the Harapan Kita Hospital in West Jakarta, Indonesia, for their kindness in providing the thalassemia dataset. Also, the authors are thankful for all reviewers involved in the process of improvement of this article.

References
[1] Galanello R and Origa R 2010 Beta-thalassemia Orphanet J. of Rare Diseases 5(11) pp 1–15
[2] Vincent O, Oluwaseyi B, James B and Saidat L 2016 Coinheritance of B-Thalassemia and Sickle Cell Anaemia in Southwestern Nigeria Ethiop J Health Sci 26(6) pp 517–522
[3] Nigam N, Nigam S, Agarwal M, and Singh P K 2017 B-thalassemia: from clinical symptoms to the management Int. J. of Contemporary Med. Research 5(5) pp 1066–1070
[4] Thalassemia International Federation https://thalassaemia.org.cy/ Accessed 23 June 2019
[5] Paokanta P, Harnpornchai N, Srichairatanakool S and Ceccarelli M 2011 The knowledge discovery of B-Thalassemia using principal components analysis: PCA and machine learning techniques Int. J. of e-Education, e-Business, e-Management and e-Learning 1(2) pp 169–174
[6] Wongseere W, Chaiyaratana N, Vichittumaros K, Winichagoon P, and Fucharoen S 2007 Thalassemia classification by neural networks and genetic programming Inf. Sci. 177 pp 771–786
[7] Yousefian F, Banirostam T and Azarkeivan A 2017 Prediction thalassemia based on artificial intelligence techniques: a survey Int. J. of Adv. Research in Comp. and Com. Eng 6(8) pp 281–287
[8] Rong Z, Yong Z, Shengzhong F and Nurbol L 2018 A novel hierarchical clustering algorithm based on density peaks for complex datasets Hindawi: Complexity 2018 pp 1–8
[9] Stitson M O, Weston J A E, Gammerman A, Vovk V, Vapnik V 1996 Theory of support vector machines. Technical Report CSD-TR-96-17, Royal Holloway, University of London
[10] Bezdek J C 1981 Pattern recognition with fuzzy objective function algorithms (New York: Plenum Press)
[11] Liu L, Shen B and Wang X 2014 Research on Kernel Function of Support Vector Machine In: Huang YM., Chao HC., Deng DJ., Park J. (eds) Advanced Technologies, Embedded and Multimedia for Human-centric Computing. Lecture Notes in Electrical Engineering (Springer, Dordrecht)
[12] Bentley J L 1975 Multidimensional binary search trees used for associative searching. Com. of the ACM 18(9) pp. 509-517
[13] Sokolova M, Japkowicz N, and Szpakowicz S 2006 Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. Adv. in Artificial Intelligence, Lecture Notes in Comp. Sci. 4304
[14] Ting K M 2011 Precision and Recall In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning (Springer, Boston, MA)
[15] Rahideh A and Shaheed M H 2011 Cancer classification using clustering-based gene selection and artificial neural networks. Proc. 2nd Int. Conf. on Control, Inst., and Automation. (Shiraz, New Jersey: IEEE) pp 1175–1180
[16] Kohavi R 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection Int. Joint Conf. on Artificial Intelligence
[17] Raju S K and Murty M R 2018 Support Vector Machine with k-fold cross validation model for software fault prediction Int. J. of Pure and Applied Maths pp 321-334