*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 2, Ver. IV (Mar-Apr. 2016), PP 86-92*
*www.iosrjournals.org*

# Analysis of C4.5 and K-Nearest Neighbor (KNN) Method on Algorithm of Clustering For Deciding Mainstay Area

Heru Ismanto[1], Retantyo Wardoyo[2]
[1]*information Systems Dept, Universitas Musamus Merauke, Papua – Indonesia*
[2]*computer Science And Electronic Dept, Universitas Gadjah Mada, Yogyakarta – Indonesia*

**Abstract :** *Development as a sustainable activity needs a good plan, so the programs can be effective and have a clear objective. Therefore, a model to help the analysis is significantly needed in determining the priority area to conduct better development in the future. This research applies the concept of Klassen Typology to analyze PDRB data in Papua Province. Based on the result of using Klassen typology analysis method, there are 4 (four) quadrants of area classification in Papua Province. Twenty nine regencies were analyzed based on PDRB data to investigate which area can be used as the development of priority area in the future. The method used in this study is C4.5 and k-nearest neighbor . Time complexity becomes test standard of a particular algorithm to get efficient execution time when implemented into programming language. The approach of asymptotic analysis using the concept of Big-o was one of the techniques that are usually used to test time complexity of an algorithm. Based on the testing result of both methods, it shows that the result of running time of KNN is more stable than of C4.5 although the analysis of Big-O gives complexity of the same time.*

***Keywords:*** Klassen Typology, C 4.5, KNN, Clustering, Priority Area

## I. Introduction

Development process is the activity done by the government for the sake of repairing and increasing the society’s living rate in order to get prosperity and high quality of life by utilizing all natural sources and human sources owned by particular areas[1]. Development of course needs good plans and clear directions, so the objectives of the development can be reached. One of development plans that can be done is to do observation of the achievement of the area which becomes the objective of the development in order to see how far development can be successfully done to the area. There are various methods used to analyze the result of development such as Klassen Typology, shift share analysis, Williamson Index, Location Quotient, Theil Index, Lorenz Curve, Ration, and L Index. Those methods are used to measure the achievement of development in a particular area based on any particular indications [2], [3]. The result of the measurement, in further, can be used as the reference to conduct more development in the future.

Previous researchers like [4] did a research about reliable area in South Kalimantan in 1993-1996. The analysis used in the research was Klassen Typology, LQ, Regional Specialization Index with multi Logit Nominal Logistic Regression Model. The variables used were PDRB, PDRB per capita, and specialization index. The research concluded that the consideration of reliable area in South Kalimantan only based on the application of per capita income and its superior sector. Besides, the growth of PDRB and area specialization did not become the most significant consideration material of determining the reliable area. [5] Did a research about reliable area in Yogyakarta Special Regency and its surrounding in 1993-1996. The analysis equipment used were Klassen Typology, LQ, Regional Specialization Index with Logit model. The variables which are used cover PDRB, per capita PDRB, and specialization index. The conclusion mentioned that only per capita PDRB variable gave significant influence to the classification of reliable area. However, the variables of PDRB growth and specialization index did not significantly influence toward the classification of reliable area. Clustering is one of the techniques of data grouping based on similarity of a particular characteristic of the data, so it can form several clusters based on the similarity of the data [6]. Several algorithms have been developed by the experts to get data clustering; some of them are partition clustering and hierarchical.

The rise of clustering algorithms becomes the interest of the experts to test those algorithms. The examples are the appearance of K-means algorithms which is believed to have the capability of faster cluster computation to various data cases. However, it has a lack because it has to be started by determining the center of the cluster before doing the process of clustering data [7], so it triggers a new variant of algorithm which cooperates the concept of fuzzy logic and K-means, and it is widely known as fuzzy c-means developed [8].

Several tests were done; they were focused on how the algorithm handles the data, working quality in processing the data, and creating the groups of the data, flexibility, and popularity of the algorithm used [9], [7]. There were only a few that lead to the test of algorithm complexity. Complexity here is meant to test and investigate the time, memory, and the other sources needed in a particular algorithm to finish computation

DOI: 10.9790/0661-1802048692

problem [10] through several approaches. One of them is asymptotic analysis by using asymptotic notation or Big-O notation for the sake of analyzing time complexity of a particular algorithm.

This research compares two methods, those are C4.5 and k-nearest neighbor. The sample used in the testing of both methods uses the data of deciding mainstay area of development in Papua Province. And then calculating the complexity of both methods.

## II. Literature Review

### 2.1 The Analysis of Lessen Typology

Lassen Typology basically divides the area based on two important indicators; those are the economic growth of the area and per capita income of the area. In this research, economic growth is measured by the growth of PDRB, while per capita income of the area is measured by per capita PDRB. The area that becomes the object of the study can be divided into four classifications; those are high growth and high income area, high income but low growth area, high growth but low income area, and low growth and low income area [11], [12].

Through these analyses, there were four classifications that could be obtained by using the approach of area and the approach of sector. Each regency/municipality which has different characteristic of economic growth and per capita income is classified by Klassen Typology with the approach of area [11].

PDRB becomes the result of additional value of the numbers of all business units owned by any particular area, and it is calculated based on the price in the market as well as constant price (Statistic Division, Real Sector Indonesian Bank, 2014). PDRB with the price in the market is used to investigate the capability of economic sources in real ways from year to year by ignoring the price factor. PDRB can be classified based on nine supporting indicators, such as: 1) agriculture, farming, forestry, and fishery, 2) mining, 3) processing industry, 4) electricity, gas, and clean water, 5) building, 6) trade, hotel, and restaurant, 7) transportation and communication, 8) finance, real estate, and company service, 9) services (including government service). In this research, the sources of PDRB Papua Province in 2012 used was based on constant price, with the assumption that the analysis result was used to analyze the level of the development based on real growth of a particular area.

### 2.2 C 4.5 Algorithm

Classification is one of techniques of collecting data. Classification (taxonomy) becomes a process of placing the objects or particular concepts into one set category based on the objects used. One of the classification techniques, which become the most popular among others, is decision tree [13]. Decision tree is one of classification methods using representation of tree structure in which every node represents the attribute; the branch represents the value of the attribute; and the leaves represent the classes. The most top node from decision tree can be called as root. There are three kinds of node in decision tree; they are:

- Root Node
  It is the highest node. In this node, there is no input, and it sometimes does not have output, or the output can be more than one.
- Internal Node
  It is branch node. In this node, there is only one input, and it has two outputs as the minimum ones.
- Leaf Node or Terminal Node
  It is the last node. There is only one input in this node, and it does not have output.

There are a lot of algorithms that can be used in forming the decision tree; those are ID3, CART, and C 4.5 [14]. The data in decision tree is usually described as table with its attribute and record. The attribute describes a particular parameter created as the criteria in forming the tree. The example is in tennis match; there are some factors that should be considered like weather, wind, and temperature.

Figure 1 shows that decision tree depends on the rule of if-then, but it does not require parameter and metric. The simple structure of decision tree can be used to solve the problem of multi-type attributes. Decision tree can also used to manage any lost values or data noise [15].

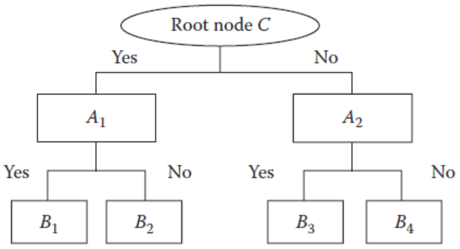

Figure 1. The Example of Decision Tree Structure

Source: Dua & Xian

Algorithm C 4.5 and decision tree are two models that cannot be separated because it needs C 4.5 algorithm to build a decision tree. In the end of 1970 until the early of 1980s, J Ross Quinlan, a researcher in learning machine, developed a model of decision tree that is called as ID3 (Iterative Dichotomiser) although the project was previously built by E. B. Hunt, J. Marin, and P. T. Stone. In further, Quinlan made an algorithm from the development of ID3 that is named as C 4.5 which bases on supervised learning [13].

A set of reparation was done to ID3. It reached the top by creating a practical system, and it is influential for decision tree which is C 4.5. The reparation covers some methods to overcome numeric attributes, missing values, noisy data, and the rules that creates rules from trees [16].

There are several steps in making a decision tree in C 4.5 algorithm [14]; those are:

1. Preparing training data. Training data is usually taken from the historical data that has ever been done in the past, and it has been classified in several classes.
2. Counting the root of the tree. The roots will be taken from the attributes that will be taken by counting gain value from each attribute. The highest gain value will become the first root. Before counting gain value of the attribute, first of all, count the entropy value. To count the entropy value, people can use the equation

Entropy (S) : $\sum_{i}^{c} pi\ log2\ pi$

Information:
C : the number of values that can be seen in the targeted attribute (the number of classification class)
Pi : the number of the samples

**2.3 K-nearest neighbor (KNN) Algorithm**

Classification is the process to find the model and function which explains or differentiate the data concept or data class for the sake of able to estimate the class of a particular objects of which label is unknown. The model itself can be in the form of the rules of “if-so” within the shape of decision tree, mathematical formula or neutral network. Classification methods are C 4.5, RainForest, Naïve Bayesian, neutral network, genetic algorithm, fuzzy, case-based reasoning, and k-Nearest Neighbor [17, [14].

K-NN algorithm is a method which uses supervised algorithm [18], [14], [13], [19]. The difference between supervised learning and unsupervised learning is on the objective of supervised learning which focuses on finding a new pattern in the data by connecting the pattern of the data that has been available to new data. On the other hand, unsupervised learning, the data has not had any pattern, and the objective of unsupervised learning is to find the pattern of the data[18], [14], [13], [19]. The objective of k-NN algorithm is to classify new objects based on the attributes and training samples [14], [13]. The result of new test samples is classified based on the majority of k-NN category. In classification process, this algorithm does not use any model to be matched, and it is only based on memory. K-NN algorithm uses neighborhood classification as the predictive value from new test sample[18], [14], [13]. The distance used is Euclidean Distance. Euclidean Distance is the most general distance used in numerical data [20]. Euclidean distance is defined as follow [13]:

$$d(xi,xj) \quad = \sum_{n=1}^{n} \left(a_r(\mathrm{xi}) - a_r(\mathrm{xj})\right)2$$

Information:
$d(x_i, x_j)$ : Distance
$x_i$ : record to i
$x_j$ : record to j
$a_r$ : data to r
i j : 1,2,3,4, …n

K-NN algorithm is the algorithm which determines the value of the distance to the test of data testing with data training based on the smallest value from the closest neighbor [20].

This part discusses the steps of the research conducted by the researcher. The research was initially started by data collecting of PDRB indicator in every regency/municipality in Papua Province from the official site of the Institution of National Statistic Center, both Province and Regency. There were nine indicators used as the data in the research; those were agriculture (V1), mining (V2), processing industry (V3), electricity and clean water (V4), building (V5), trade-hotel and restaurant (V6), transportation and communication (V7), finance-renting and company service (V8), and other services (V9). Twenty nine regencies were analyzed in this research by using Klassen Typology to determine reliable area in Papua Province. In this research, Province and Regency/Municipality PDRB year 2012 and economic growth data from 2006 until 2012 were used. The analysis result is used to create running time analysis by using KNN and C 4.5 methods to get complexity value, and then the data from both methods was compared. In further, the algorithm was tested by using Weka application toward several data sets of the testing to see running time needed in every algorithm. The test was done nine times by dividing data sets into several parts. Every test was done three times measurement, so the total of the test was twenty-seven times (27x). The test result of running time was analyzed to obtain the information of correlation between Big-O analysis and running time in every algorithm.

## III. Result And Discussion

### 3.1 The Analysis of Complexity Comparison

This section talks about how every procedure in algorithm which counts time complexity by using Big-O. the discussion of algorithm in this research consists of two kinds of algorithms; those are: C 4.5 and k-nearest neighbor (k-NN or KNN). To obtain the value of complexity from both algorithms, there should be pseudo code from the algorithm created to ease the implementation into programming language or within the programming tools. Here it is the discussion about complexity from each algorithm by taking some part of the core of pseudo code in order to count the complexity.

### 3.1.1 The Classification Test of Reliable Area

In conducting running time test, sample data was used from the result of Klassen Typology analysis forming four quadrants of areas in Papua Province; those are: time series data of economic growth occurrences in the whole regencies/municipalities from 2006 until 2012. It consists of nine variables and two thousand (2,000) data sets. The test was conducted by giving the number of data sets such as five hundred (500), a thousand (1,000), and two thousand (2,000) which became nine indicators from PDRB data; twenty-nine (29) regency/municipality areas as well as economic growth which was counted from 2006 until 2012. Every test was divided into three measurements of running time to make sure the accuracy of running time owned by each algorithm when different data sets were given.

In the process of making data classification model, the algorithms used were C 4.5 and k-nearest neighbor (k-NN or KNN). The reference of source code was taken from Waikato University, Hamilton, New Zealand, as the party of Weka Tools, and it would be used as assisting tools for forecasting the measurement.

### 3.1.2 C 4.5 Algorithm

Here it is the division of the core of pseudo code from C 4.5 algorithm:

```
Algorithm C4.5(D)
Input: an attribute-valued dataset D
  1: Tree={}
  2: if D is" OR other stopping criteria met then
  3:   terminate
  4: end if
  5: for all attribute α Є D do
  6:   Compute information-theoretic criteria if we split on α
  7: end for
  8: α_best = Best attribute according to above computed
     criteria
  9: Tree = Create a decision node that tests α_best in the root
 10: D_v = Induced sub-datasets from D based on α_best
 11: for all D_v do
 12:   Treev = C4.5(Dv)
 13:   Attack Tree_v to the corresponding branch of Tree
 14: end for
 15: return Tree
```

Figure 2. Some Part of Pseudo Code C 4.5 Algorithm which was Counted its Complexity

By referring to Figure 2, line (1) shows the process of choosing initial data input from the group of data set D. at least this line needs a random search within n times, so time complexity owned by line (1) was 0 (1).

Every time line (2) was running, so at least once repetition would be operated within it; that was line (5) and line (7). Similar case also works for every repetition that exists in line (11). The repetition within it would run n times in every repetition body, so the repetition in line (3) had complexity with O (1), it is also similar to the repetition in line (11) which had time complexity O (1).

Based on the algorithm above and the figure, it can be identified that C 4.5 Algorithm had complexity Big-O (1).

### 3.1.3 K-nearest neighbor (KNN) Algorithm

```
1:  BEGIN
2:  Input x, of unknown classification
3:  Set K, 1 ≤ n
4:   Initiaze i = 1
5:  DO UNTIL (K-nearest neighbors to x found)
6:  Compute distance from x to $x_i$
7:  IF (i ≤ K) THEN
8:  Include $x_i$ in the set of K-nearest
9:  neighbors
10:  ELSE IF ($x_i$ closer to x than any
11:  previus nearest neigbors) THEN
12:  Delete the farthest of the K-
13:  nearest neighbors
14:  include $x_i$ in the set of K-
15:  nearest neighbors
16:  END IF
17:  END DO UNTIL
18:  Initialize i = 1
19:  DO UNTIL (x assigned membership in all
20:  classes) Compute $u_i(x)$ using (1)
21:  Increament i
22:  END DO UNTIL
23:  END
```

Figure 3. Some Part of Pseudo Cade of k-nearest neighbor with Counted Complexity

Pseudo code in Figure 3 shows that line-2 until line-4 had time complexity within O(1); line 5 until line 17 had complexity within O(n). on the other hand, line 18 was O (l), and line 19 until line 22 was O(n). Thus, Total complexity = O(1) + O(n) + O(l) + O(n) = O(n)

3.2 The Analysis of Running Time

This part discusses about the implementation of every algorithm into Weka application to test the running time owned by every algorithm. The test was divided into three steps; those were by cutting the data set within 500, 1,000, and 2,000. Every test was determined into three times running time measurement to make sure running time accuracy owned by every algorithm to handle several different data inputs. Here they are the tables of testing result that had been done. Time measurement was done based on second unit (00).

Table 1. First Measurement

| Alg. | Sample Data set | | |
|---|---|---|---|
| | 500 | 1000 | 2000 |
| C45 | 0.01 | 0.02 | 0.03 |
| K-NN | 0.21 | 0.51 | 0.84 |

Table 2. Second Measurement

| **Alg.** | **Sample Data set** | | |
|---|---|---|---|
| | **500** | **1000** | **2000** |
| **C45** | 0.01 | 0.03 | 0.04 |
| **K-NN** | 0.20 | 0.52 | 0.83 |

Table 3. Third Measurement

| **Alg.** | **Sample Data set** | | |
|---|---|---|---|
| | **500** | **1000** | **2000** |
| **C45** | 0.01 | 0.01 | 0.03 |
| **K-NN** | 0.21 | 0.50 | 0.82 |

Table 1-3 becomes the result of measurement toward all tests to all data sets that had been determined into several parts. C 4.5 Algorithm in the first, second, and third measurements had increasing running time every test by having different data set numbers. The more data set, the bigger running time needed by C 4.5 Algorithm compared to KNN Algorithm. It also works for the second and third measurement of C 4.5 Algorithm. (See Figure 4-6 for running time comparison graphs in every algorithm).

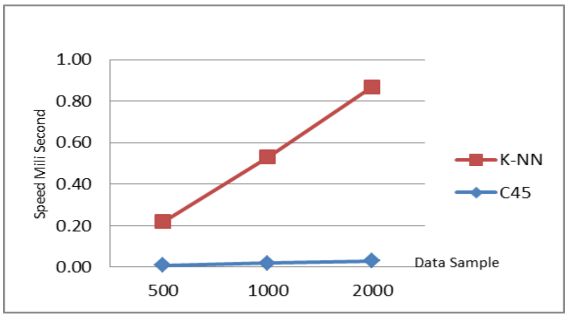

Figure 4. The Graph of The First Running Time Measurement

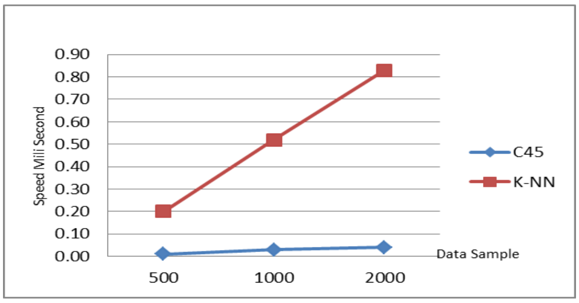

Figure 5. The Graph of The Second Running Time Measurement

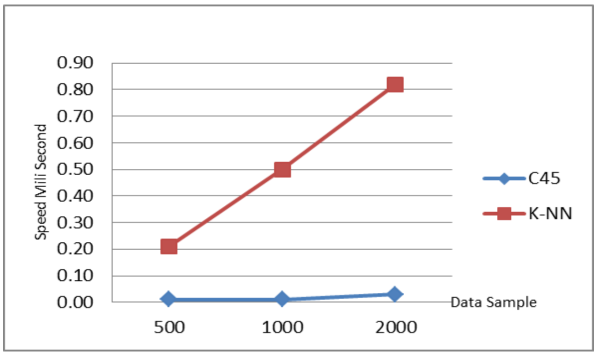

Figure 6. The Graph of The Third Running Time Measurement

On the other hand, KNN Algorithm from the first, second, and third measurements (see table 1-3) needed running that was needed, and the character was linear and global. The increase need of running time for big number of data sets had significant influence to the use of big running time. (See Figure 5 for running time comparison graph to the second measurement).

However, both algorithms had running time which did not really have big difference toward the data presented.

## IV. Conclusion

Based on the result of analysis and testing using three trials that running time is greatly affected by the complexity of algorithm used and the increasing number of set data as the input. As the testing result has been applied, it shows that method of C4.5 is more stable than of k-nearest neighbor. Furthermore, observing from the result of three trials above it can be concluded that running time is greatly affected by the complexity of algorithm and the increasing number of set data as the input. As the trial result above, it can be analyzed that the running time of two algorithms in a series are C4.5 and then KNN

## References

[1]. Mursidah, A. Hamzah, dan Sofyan, Analisis Pengembangan Kawasan Andalan Di Kabupaten Aceh Besar, Jurnal Ilmu Ekonomi, Volume 1, No. 1, Februari 2013, pp. 43- 55, 2013.

[2]. Yunisti, Trias Dewi, Analisis Ketimpangan Pembangunan Antar Kabupaten/Kota Di Provinsi Banten, Tesis Program Magister, 2012

[3]. Alfita, Riza, Perancangan Sistem Pendukung Keputusan Penentuan Prioritas Produk Unggulan Daerah Menggunakan Metode Weighted Product (WP), Prosiding Seminas Competitive Advantage, Vol 1, No 1 (2011), 2011.

[4]. Dhyatmika, Ketut Wahyu, Analisis Ketimpangan Pembangunan Provinsi Banten Pasca Pemekaran, Skripsi Fakultas Ekonomi dan Bisnis Universitas Diponegori, Semarang, 2013

[5]. Aswandi dan M. Kuncoro, Evaluasi Penetapan Kawasan Andalan : Studi Empiris di Kalimantan Selatan 1993 – 1999. Jurnal Ekonomi dan Bisnis Indonesia. Vol 17, No. 1, hal. 27-45, 2002.

[6]. KusumaDewi, Analisis Kawasan Andalan Sebagai Pusat Pertumbuhan dan Penggerak Perekonomian di Propinsi DIY dan Sekitarnya, Skripsi FE UGM, 2002.

[7]. Witten, H. Ian dan Eibe Frank, Data Mining : Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann Publishers, San Fransisco, 2005.

[8]. Alfina, Tahta, S. dan A. R. Barakhah, Analisa Perbandingan Metode Hierarchical Clustering, K-means dan Gabungan Keduanya Dalam Cluster Data (Study Kasus : Problem Kerja Prakterk Jurusan Teknik Industri ITS), Jurnal TEKNIK ITS, Vol. 1, pp. 521 – 525, 2012

[9]. Ross, J. Timothy, Fuzzy Logic With Engineering Applications 3rd Edition, John Wiley and Sons Ltd. Publication, Mexico, United State of Americca, 2010.

[10]. Abu Abbas, Osama., Comparisons Between Data Clustering Algorithms, The International Arab Journal of Information Technology, Vol. 5, No. 3, pp. 320 – 325, 2008.

[11]. Sipser, Michael, Introduction to the Theory of Computation – Second Edition, Thomson Course Technology, Massachusetts, 2006.

[12]. Syafrizal, Pertumbuhan Ekonomi dan Ketimpangan Regional Wilayah Indonesia Bagian Barat. Prisma. LP3ES. Tahun XXVI hal. 27–38, 1997.

[13]. H. Hill (Ed), Unity and Diversity : Regional Economic Development in Indonesia since 1970, Oxfort : University Press, 1989.

[14]. J. Han, & M. Kamber, Data Mining Concept and Tehniques, San Fransisco, Morgan Kauffman. ISBN 13: 978-1-55860-901-3, 2006.

[15]. D. T. Larose, Discovering Knowledge in Data. New Jersey, John Willey & Sons, Inc. ISBN 0-471-66657-2, 2005.

[16]. S. Dua, & Du. Xian, Data Mining and Machine Learning in Cybersecurity, USA, Taylor & Francis Group, ISBN-13: 978-1-4398-3943-0, 2011.

[17]. I. H. Witten, E. Frank, M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques (3rd ed). USA, Elsevier, 2011.

[18]. A.S. Arriawati, Y. Chrystiyono, Klasifikasi Citra Tekstur Menggunakan k-Neirest Neighbor Berdasarkan Ekstraksi Ciri Metode Matriks Kookurensi, Available from: Http://jurnal pointer. ub.ac.id/index.php/pointer/ article/view/11, 2011.

[19]. X. Wu, V, Kumar, The Top Ten Algorithms in Data Mining, New York, CRC Press, 2009.

[20]. T. Mitsa, Temporal Data Mining, New York , CRC Press, 2010.

[21]. G. Goujon, Chaoqun, W. Jianhong, Data Clusterin :Theory, Algorithms, and Applications, Virginia, ASA, 2007.