A Modified K-Means with Naïve Bayes (KMNB) Algorithm for Breast Cancer Classification

Dian Eka Ratnawati[1], Nurizal Dwi Priandani[1], Machsus[2]

[1]*Department of Computer Science – Faculty of Computer Science, Brawijaya University.*

[2]*Department of Civil Infrastructure Engineering, Institute of Technology Sepuluh Nopember.*

dian_ilkom@ub.ac.id

***Abstract*—Breast cancer is a second biggest cause of human death on women. The death rate caused by the breast cancer has been fallen since 1989. This downfall is believed as a result from early diagnose on breast cancer, the awareness uplift on the breast cancer, also a better medical treatment. This research proposes the Modified K-Means Naïve Bayes (KMNB) method on Breast Cancer data. The modification which has been conducted was an additional on initial centroid which has been proposed by Fang. The experiment compared the accuracy of our proposed method with the original KMNB, Original K-Mean, and K-Means using initial centroid by Fang. Based on the result of the experiment, the accuracy of our proposed method was 95%. The error reduction of our proposed method was about 50% compared to the original KMNB. It can be stated that our proposed method is promising and able to enhance the prediction on Breast Cancer Wisconsin data. On the other hand, the enhancement of prediction result will increase the preventive behavior on society and give a positive impact on the number downfall of breast cancer sufferers.**

***Index Terms*—Breast Cancer; Classification; Clustering; Data Mining.**

I. Introduction

Breast cancer is breast cells which grow uncontrollably. The "Breast Cancer" term is a reference for a malignant tumor which grows from the breast cells [1]. The tumor would be cancerous when the cells able to grow and invade the tissue or metastasis to the area around the body. Breast cancer commonly occurs for a woman, but there is a probability for man [2].

Breast cancer is a second biggest cause of human death on a woman. The probability of woman death caused by the breast cancer is 1:36. The deathrate caused by the breast cancer has been fallen since 1989, mainly on under 50 years woman. This downfall is believed as a result from early diagnose on breast cancer, the awareness uplift on the breast cancer, also a better medical treatment [3].

Generally, there are two methods to predict on a dataset, clustering, and classification. Clustering is a process of dividing the data into groups which have identic object characteristically. [4]. Clustering is a process of item data grouping into a number of a small group until each group has an essential similarity [5]. Cluster analysis [6] is a major method for data analysis which has broad function and practical applications in emerging areas like Bioinformatics [7, 8]. Classification is a process to find models (or function) which describe and able to differentiate concepts or classes of data. The purpose of the classification is to make the model able to be used for predicting the class of any object or data which has an unknown class label [9].

K-Means algorithm is a most popular clustering algorithm and many used for industry. K-means conventional is arranged based on a simple idea. At first, the amount of the cluster must be determined. Any objects or the first element in the cluster is able to be chosen as a centroid point cluster [10]. K-Means Algorithm [6, 11, 12, 13, 14] is effective in producing the cluster for many practical applications, but the complexity of the k-means algorithm computation is immense, mainly on a huge dataset. Furthermore, the algorithm result on various cluster depends on the early random choice of the centroid. Some researches have been conducted by researchers to enhance the clustering k-means algorithm performance. Two of Modified K-Means are K-Means Naïve Bayes (KMNB) by Meiping [15] and K-Means with initial centroid by Fang [16].

On this research, the modified KMNB will be conducted by combining the KMNB which have been done by Meiping and adding the initial centroid which has been done by Fang. The Novelty of this research is the hybridization of the initial centroid method Fang on the KMNB early process, on the breast cancer study case. The modified KMNB with initial centroid by Fang is expected to be able to enhance the accuracy of the breast cancer data classification process.

II. Related Work

K-means algorithm is a most studied and implemented on study case clustering algorithm, some on paper [17] and [18]. On paper [17], K-means has been implemented on Network Flow Classification study case. On paper [18], K-means has been implemented on Remote Sensing Image Classification. Generally, K-means algorithm can produce a better result on these studies. The major weakness of original K-means [19] is producing a different cluster on different initial centroid value set. As result, the quality of the cluster is very dependent on the choice of the initial centroid. K-means algorithm is an expensive computation and requires time which proportional with the product from the amount of the item data, the amount of cluster and iteration. Based on that limitation, many research have been resulted to optimize the performance of K-Means.

Some efforts have been conducted by researchers to enhance the effectivity and efficiency of K-means algorithm [16, 18, 19, 20]. K-mode is a variant based on K-means algorithm [21] by replacing the cluster average with modes. K-modes algorithm produces the Local optimum solution which depends on the choice of initial mode. K-prototype algorithm [22] integrates K-means and processing the K-modes for data grouping. In this method, the measurement of dissimilarity is defined by taking into account both numeric and categorical attributes.

Fahim et al. [20] proposed an efficient method to determine the data point for a cluster. Original K-means is an expensive computation which expensive because every iteration counts the range between a data point and every centroid. The approach by Fahim used two range function, for this purpose, similar to the K-means and the another one based on heuristic to decrease the amount of range calculation, but this method assumes that the initial centroid is randomly chosen, such as the original K-means algorithm case. Therefore, there is no guarantee for the accuracy of the final cluster.

Some methods related to the random selection problem of a number of clusters and initialization of centroids have been proposed. On the research [18] have been combining the K-Means with hill climbing to solve this problem. Fang et al. [16] proposed a more systematic method to find the initial centroid. Centroid which obtained by this method was consistent with the data distribution. Therefore, the resulted cluster had a better accuracy compared to the original K-means algorithm.

Some earlier researches have been proposed to enhance the K-means algorithm accuracy with the hybridization of K-means with another algorithm. The research which has been conducted by Meiping [15] and Muda [23] discussed the integration of K-means with Naive Bayes or usually called as KMNB. Meiping [15] used KMNB to classified the problem risk of bank credit card. The result of the research concluded that the integration method of Naive Bayes and K-Means have a better accuracy than Bayesian method. The other research which related with KMNB which conducted by Muda [23], had an accuracy result of KMNB on dataset KDD Cup 1999 reach 99,6% when used the naïve Bayes algorithm only, the accuracy was 83.19%.

III. PROPOSED METHOD

On the improved KMNB algorithm which discussed in this paper, the main process on original KMNB based on the [15] was modified. The modification was a method change to configure the initial centroid. The modification purpose was to the accuracy. The proposed flowchart of modified KMNB shown in Figure 1.

In the first phase, clustering process using K-Means algorithm was conducted first. On our proposed method, the configuration of K-Means initial centroid was not random. The determination of initial centroid was conducted systematically by using the method which proposed by Fang on [16]. Breast-Cancer-Wisconsin Dataset [24] basically had two classes, benign and malignant. In the first phase, K-Means has used to grouping the data based on three clusters which are, malignant, benign, and "Maybe" (does not included in a malign/benign cluster). The prediction class on malignant and benign was gathered from the malignant and benign clusters representation which resulted from the first phase. The "Maybe" cluster which resulted from the first phase was used in the second phase. Figure 2 shows the illustration of the first phase result.

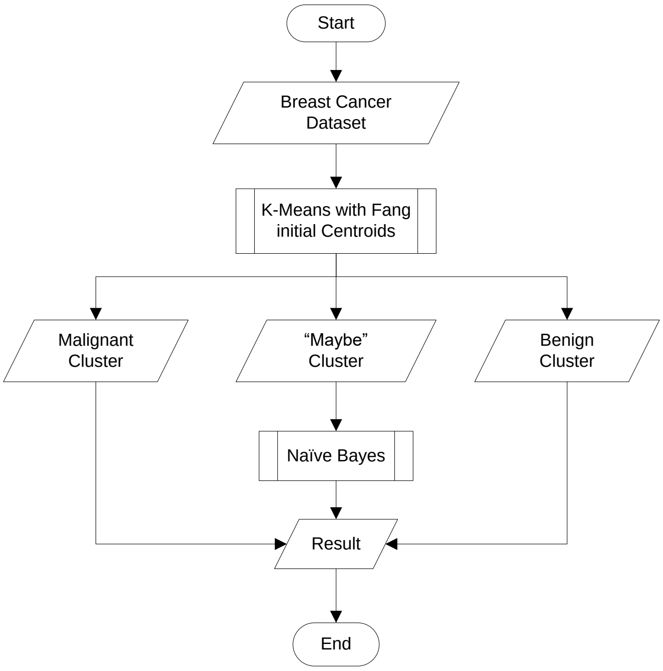

Figure 1: Flowchart of our proposed method

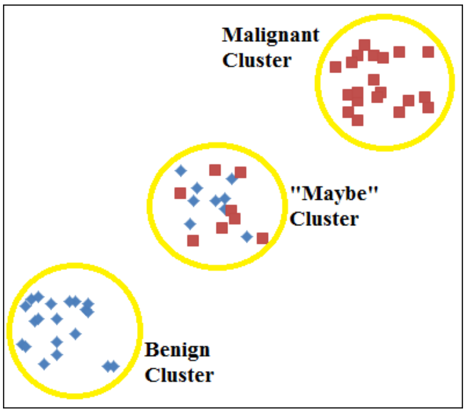

Figure 2: Illustration of the first phase result

Naïve Bayes Classification algorithm was used in the second phase. The "Maybe" cluster data have been classified into a malignant or benign category in this phase. As shown in Figure 2, the data of “Maybe” cluster consisted on benign and malignant data class. Training and testing data which have been used in the second phase was a data which can be found in the cluster itself. Every data have been classified into benign or malignant classes as the final result in this phase.

IV. EXPERIMENT AND DISCUSSION

A. *Experiment Setup*

The accuracy result based on our proposed method, compared to the original KMNB, Original K-Means, and K-Means using initial centroid by Fang (K-Means + Fang), have been used as the benchmark. Breast Cancer Wisconsin Dataset from UCI repository [24], have been used in this algorithm as a test case. The data amount which used in this research were 402 data record. The dataset was divided into two randomly as the training data and testing data. The training data variation which used were 70% (475), 60% (408), and 50% (340). Each test case has been conducted multiple times, as many as five times to gathered the average result.

B. Experiment Result and Discussion

The experiment result gathered from each test case shown in Table 1. To visualize the experiment result, a graphic have been made as shown in Figure 3.

Table 1
Experiment Results

Training Data Presentation	Original K-Means	(K-Means + Fang)	Original KMNB	Our Proposed Method
50%	91%	92%	93%	95%
60%	90%	92%	93%	95%
70%	91%	92%	94%	95%

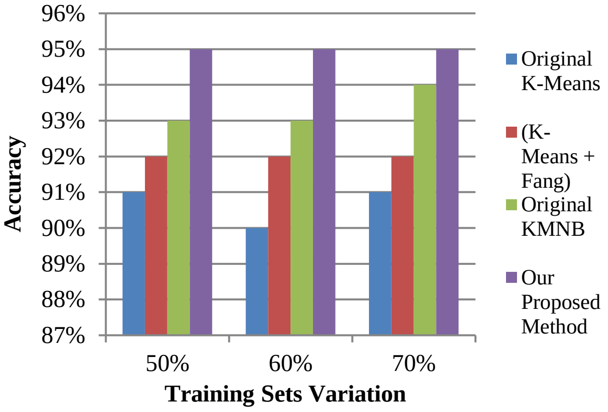

Figure 3: Graphic of experiment result

Based on Table 1 which visualized on the graphic in Figure 3, it is shown that the accuracy result of our proposed method was better compared to original KMNB, Original K-Means or K-Means using initial centroid by Fang. Our proposed method gave 95% accuracy on every training data variation. Compared to original K-Means, KMNB, and K-Means using initial centroid by Fang, our proposed method increased about 5% accuracy. Nevertheless, our proposed method was able to reduce the error rate up to 50%.

The cluster which produced in the first phase was more promising after the addition of the initial centroid method by Fang. It was caused by our proposed method which has an exact scheme on the initial centroid determination which has been conducted by Fang. In [16], the proposed initial centroid determination method result for K-Means has been proved more stable compared to randomize initial centroid on original K-Means. Furthermore, the utilization of Naïve Bayes Classifier on the second phase has a role in improving the accuracy by classifying on the "Maybe" cluster. The "Maybe" cluster would make a negative impact on the accuracy when it was not conducted. The reason was because the cluster became outlier or noise data.

The first phase generated the majority data which have been clustered based on its original class, malignant and benign. As result, data in the probable cluster was lesser. The data prediction on “maybe” cluster have been repaired by using Naïve Bayes Classification in the second phase. Related to the amount of “maybe” cluster data which inconsiderable, the accuracy result was increased although insignificant. It was the impact of Naïve Bayes classification that required enough training data to conduct an optimal classification process.

Different from our proposed method, original K-means generated initial centroid randomly which made the accuracy quality of clustering was depended on its initial centroid. When the centroid was not precise, the accuracy result would be relatively lower. On K-Means with the initial centroid by Fang, the accuracy result was more stable. It was affected by the initial centroid determination which more definite than centroid choice randomly. In original KMNB, the initial centroid determination K-Means process also affected the accuracy. However, the impact was able to be reduced by the addition of the Naïve Bayes Classifier in the second phase which resulted in a better accuracy, although was not better than our proposed method.

V. CONCLUSION

This paper proposed an improved K-Means Naïve Bayes (KMNB) method on Breast Cancer data. The modification which has been conducted was an additional on initial centroid which has been proposed by Fang. The accuracy of our proposed method was about 95%. Compared to original KMNB, Original K-Mean, and K-Means using initial centroid by Fang, our proposed method gave a better result. The addition of Fang's method to determine the initial centroid was critical to the accuracy quality improvement compared to another method which has been mentioned in this paper.The prove was on error reduction of our proposed method, it was about 50%. It can be stated that our proposed method is promising and able to enhance prediction on data Breast Cancer Wisconsin. On another hand, the increase on prediction result will increase a preventive behavior in the society and reduce the number of Breast Cancer sufferers.

The initial centroid increase by Fang on KMNB had a positive impact which is increasing the prediction accuracy, but on the other hand, it took a longer computation time. Based on the weakness, in the future, it will need a new schema to improve the computation time without reducing the prediction accuracy result..

REFERENCES

[1] Breastcancer.org, “What Is Breast Cancer?” [Online]. Available: http://www.breastcancer.org/symptoms/understand_bc/what_is_bc. [Accessed: 21-Dec-2016].

[2] American Cancer Society, “What is breast cancer?” [Online]. Available: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-what-is-breast-cancer. [Accessed: 21-Dec-2016].

[3] American Cancer Society, “What are the key statistics about breast cancer?” [Online]. Available: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics. [Accessed: 21-Dec-2016].

[4] M. N. Murty and V. S. Devi, Pattern Recognition: An algorithmic approach. Springer, 2011.

[5] A. R. Webb and K. D. Copsey, Statistical Pattern Recognition, Third., A John Wiley & Sons, Ltd, 2011.

[6] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006.

[7] Hasan, Mohammad Shabbir, and Zhong-Hui Duan. "Hierarchical k-Means: A Hybrid Clustering Algorithm and Its Application to Study Gene Expression in Lung Adenocarcinoma." (2015).

[8] Daxin Jiang, Chum Tong and Aidong Zhang, “Cluster Analysis for Gene Expression Data,” IEEE Transactions on Data and Knowledge Engineering, 16(11): 1370-1386, 2004.

[9] Han J dan Kamber M. 2001. Data mining Concepts & Techniques.

[10] G. Dougherty, Pattern Recognition and Classification: An Introduction. Springer, 2012.

[11] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006.

[12] McQueen J, “Some methods for classification and analysis of multivariate observations,” Proc. 5th Berkeley Symp. Math. Statist. Prob., (1):281–297, 1967.

[13] Pang-Ning Tan, Michael Steinback and Vipin Kumar, Introduction to Data Mining, Pearson Education, 2007.

[14] Stuart P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, 28(2): 129-136.

[15] X. Meiping, "Application of Bayesian Rules Based on Improved K-Means Classification on Credit Card," *2009 Int. Conf. Web Inf. Syst. Min.*, pp. 13–16, 2009.

[16] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. Of the 3rd International Conference on Machine Learning and Cybernetics, pages 26–29, August 2004.

[17] A. Munther, R. Razif, M. AbuAlhaj, M. Anbar, and S. Nizam, "A Preliminary Performance Evaluation of K-means, KNN and EM Unsupervised Machine Learning Methods for Network Flow Classification," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 2, p. 778, 2016.

[18] B. S. Chandana, K. Srinivas, and R. K. Kumar, "Clustering Algorithm Combined with Hill Climbing for Classification of Remote Sensing Image," *Ijece*, vol. 4, no. 6, pp. 923–930, 2014.

[19] K. Nazeer, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," Proc. World Congr., vol. I, pp. 1–5, 2009.

[20] Fahim A.M, Salem A. M, Torkey A and Ramadan M. A, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7):1626–1633, 2006.

[21] Chaturvedi J. C. A, Green P, "K-modes clustering," J. Classification, (18):35–55, 2001.

[22] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, (2):283–304, 1998.

[23] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir , "Intrusion Detection based on k-Means Clustering and Naïve Bayes Classification", 7th International Conference on IT in Asia (CITA), 2011.

[24] Matjaz Zwitter & Milan Soklic (physicians) (1988-07-11). UCI repository of machine learning databases.X. S. Li*, et al.*, "Analysis and Simplification of Three-Dimensional Space Vector PWM for Three-Phase Four-Leg Inverters," *IEEE Transactions on Industrial Electronics*, vol. 58, pp. 450-464, Feb 2011.