Possibilistics C-Means (PCM) Algorithm for the Hepatocellular Carcinoma (HCC) Classification

IOP Conf. Series: Materials Science and Engineering **546** (2019) 052038 doi:10.1088/1757-899X/546/5/052038

Possibilistics C-Means (PCM) Algorithm for the Hepatocellular Carcinoma (HCC) Classification

Rafiqatul Khairi[1], Zuherman Rustam[1*], Suarsih Utama1

[1]Department of Mathematics, University of Indonesia, 16424 Depok, Indonesia

*Corresponding author email: rustam@ui.ac.id

Abstract. Hepatocellular Carcinoma (HCC) is a malignant tumor that attacks the liver and can cause death. Although there have been advances in technology for the prevention, diagnosis, and treatment, the number of liver cancer patients is still increasing. The liver can still function normally even if some of its parts are not in good condition. Therefore, the symptoms of liver cancer at an early stage are difficult to detect. Early diagnosis of this disease will increase the chances of recovery. One method to diagnose Hepatocellular Carcinoma (HCC) is to check the level of alpha-fetoprotein (AFP) in the blood which is alpha-fetoprotein (AFP) is a cancer index. If the liver cancer cells continue to grow, the level of alpha-fetoprotein (AFP) will be very high. This paper presents a Possibilistic C-Means (PCM) algorithm, which used to classify the results of alpha-fetoprotein (AFP) blood tests to determine whether patients diagnosed with Hepatocellular Carcinoma (HCC) or normal patients. This method will help to get an accuracy of about 92%.

1. Introduction

Cancer is the second leading cause of death in the world, with an estimated 9.6 million deaths has occurred in 2018. In general, approximately one of six deaths caused by cancer. About 70% of deaths due to cancer occur in low and middle-income countries [1]. Cancer is a disease that can grow and develop in various organs in the body. Cancer occurs when cells grow abnormally and disrupt normal cell function so that the body has difficulty working properly [2].

Cancer that grows and develops in the liver is called liver cancer. The liver is an organ located precisely below the right lung. The liver is composed of cells called hepatocytes. Hepatocytes may form some types of malignant (cancerous) and benign (non-cancerous) tumors. Some of the important functions of the liver is to store excess nutrients and return it to the bloodstream, to produce a blood protein that helps coagulation (blood clotting process to avoid heavy bleeding when injured), to transport oxygen and to help immune system function, to produce bile and deliver it to the intestines in order to help digest food and absorb nutrients, and to rid the body of harmful substances in the blood, including drugs and alcohol [3]. The liver can still function normally even if some of its parts are not in good condition. Therefore, the symptoms of liver cancer at an early stage are difficult to detect.

Hepatocellular Carcinoma (HCC) is a disease that ranks sixth in the world among a wide variety of malignant diseases and is the third leading cause of death due to cancer [4, 5]. HCC, a form of primary liver cancer, is the most common disease which very often experienced by adults [6]. Patients with HCC continue to increase worldwide. Usually, the biggest cause of HCC is chronic hepatitis B. While in Southern Europe and North America, the biggest cause of HCC is chronic hepatitis C [4]. Factors

that increase the risk of developing HCC disease are male gender, advanced age, obesity, alcohol use, diabetes, and genes factor (heredity) [4,7,8]. Treatment for patients with HCC can be adjusted to the condition of the patient's liver and their overall health. Some treatments for patients with HCC are surgery, liver transplantation, destroying cancer cells, chemotherapy or radiation directly to the cancer cells, and therapeutic use of drugs [9].

One method to diagnose Hepatocellular Carcinoma (HCC) is to check the level of alpha-fetoprotein (AFP) in the blood, which is alpha-fetoprotein (AFP), is a cancer index. If the level of AFP is high, it means that the liver cancer cells continue to grow. Usually, the level of AFP is high in the fetal blood, but not long after birth the AFP level will drop and become lower than the normal level. Furthermore, the AFP test can be used to determine what treatment may be performed on patients. Then, at the time of the treatment, the AFP test can also be used to provide an overview of the development of the patient's health. Finally, the AFP test can also be used after the treatment to see whether cancer has recurred [10].

In this paper, we use a data set from the alpha-fetoprotein blood test in the Al Islam Bandung Hospital laboratory, Bandung, West Java. This data will be classified into two parts those are data of patients diagnosed with HCC and data of normal patients. The classification of the data into two parts uses a Possibilistic C-Means (PCM) algorithm that is proposed by Krishnapuram and Keller [11]. This approach is partitioned possibilistic, where the characteristic level of a node in a cluster is measured by possibilistic membership. Membership possibilistic are rational and have a good interpretation. In the possibilistic approach, the noisy point which position is far will belong to the cluster that has a little possibilistic membership, so there is no significant effect in the clustering result. Therefore, we can say this approach is a powerful approach [12]. Several studies on the classification of Hepatocellular Carcinoma have been done using various methods, including Barcelona Classification [13], Okuda, Barcelona Clinic Liver Cancer, Cancer of the Liver Italian Program, and Japan Integrated Staging [14], Voting Ranking Random Forests [15].

2. Method

2.1. Clustering

Clustering is the technique of grouping data or object that is done by searching for characteristic similarity based on shared attributes of data or object to the same cluster [16, 17]. Clustering has been widely used for bioinformatics, data mining, computer vision, and other applied fields. Clustering aims to group a number of objects so that the objects that are in the same group have a high degree of similarity and objects in different groups have a high level of inequality [18, 19].

2.2. Possibilistic C-Means (PCM)

As an important data mining tool, Possibilistic C-Means (PCM) algorithms have emerged as an important technique for pattern recognition and data analysis, proposed by Krishnapuram and Keller [11, 20]. PCM can be used widely in the big sensor is a data analysis and data mining. However, in the big sensor data, a lot of data collection experience incompleteness; that is, the X data set can contain a vector that is missing one or more attribute values [21]. PCM cannot succeed fully in classifying a collection of incomplete data in real time. On the one hand, PCM cannot calculate the distance between two objects in a collection of incomplete data. While the accuracy of PCM can be easily damaged by objects that are not complete. On the other hand, PCM is difficult to meet the requirements of real-time sensors big grouping incomplete data because of the large amount of data. However, the advantages of PCM are there the possibility of grouping the noisy data samples can be done. Noisy data is a number of data that have noise points and outliers. As a result, inter-cluster with one another is independent. In general, PCM minimizes the following objective function [20]:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}\left(u_{ij}\right)^m \|x_k - v_i\|^2 + \sum_{i=1}^{c}\eta_i \sum_{j=1}^{n}(1-u_{ij})^m \quad (1)$$

Where:
n = number of data
c = number of cluster
V = (v_1, v_2, ...,v_c) is C-tuple of prototypes
U = cxn sized matrix called the matrix possibilistic c-partition
u_{ij} = elements of the U matrix
m = the degree of Fuzzy's, where m> 1
η_i = a positive constant, where i = 1, 2, ..., c

PCM algorithm can be described as follows [20]:
Step 1. Choose m, c, and $\varepsilon > 0$, then initialize the membership matrix $U^{(0)}$.
Step 2. Update the cluster center using the following formula:

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m} \quad (2)$$

Step 3. Estimate η_i using the following formula:

$$\eta_i = \frac{\sum_{j=1}^{n}(u_{ij})^m (d_{ij})^2}{\sum_{j=1}^{n}(u_{ij})^m} \quad (3)$$

Step 4. Update the membership of U matrix using the following formula:

$$u_{ij} = \frac{1}{1+\left(d_{ij}/\eta_i\right)^{1/(m-1)}} \quad (4)$$

Step5. If $\left\|u_{ij} - u'_{ij}\right\|^2 \leq \varepsilon$, stopped; else go to **Step 2**.

PCM are rational and have a meaningful interpretation. Even more important is the approach PCM is also strong, because of noisy objects will belong to the group with a small membership, and consequently cannot damage the group generated significantly [12]. In the PCM algorithm, each object is considered equally important in the clustering solution [20].

2.3. Confusion Matrix

In machine learning, the confusion matrix, also known as the error matrix, is a four-cell contingency table, usually used to visualize the performance of a classification algorithm. This contingency table consists of two dimensions (actual and predicted), where the predicted class is represented by each matrix line and the actual class is represented by each matrix column, or vice versa [22] (See **Table 1**).

Table 1. Confusion Matrix

		True Condition (Actual Class)	
		Positive Condition	Negative Condition
Predicted Condition (Predicted Class)	Positive Condition	**True Positive (TP)**	**False Positive (FP)**
	Negative Condition	**False Negative (FN)**	**True Negative (TN)**

Based on this confusion matrix, if the object is positive and is classified as positive, it is called a true positive; if it is classified as negative, it is called a false negative. Then, if the object is negative and is classified as negative, then it is called a true negative; if it is classified as positive, it is called a false positive [23].

The following equations are some of the calculations that we can get from the confusion matrix; accuracy, sensitivity, precision and F1Score [23]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{8}$$

Accuracy is the weighted average arithmetic of sensitivity and precision with the opposite of each other. Sensitivity (or also called recall) is the proportion of real positive cases that are predicted positive, where the desired feature is to describe how many of the relevant cases are taken by predicted positive rules. Precision is the proportion of predicted positive cases that are indeed real positives or vice versa. F1Score is the harmonic mean that combines sensitivity & precision [22].

3. Experiment

In this paper, we use a dataset of alpha-fetoprotein blood tests in the laboratory of Al Islam Bandung Hospital, Bandung, West Java (See **Table 2**).

Table 2. Dataset samples from the alpha-fetoprotein blood test in the laboratory of Al Islam Bandung Hospital, Bandung, West Java.

No.	Gender	Age (y.o.)	AFP (ng/mL)	Hemoglobin (g/dL)	Leukocytes (cell/uL)	Hematocrit (%)	Platelets (cells/uL)	Diagnosis
1	1	58	123 147	6.8	11700	21.6	150000	1
2	1	53	20	9.4	16100	28.3	476000	1
3	1	61	278.8	11.6	12300	34.4	161 000	1
4	1	64	8.1	11.3	6500	33.2	98000	1
5	2	37	1	11.9	4500	34.4	201 000	0
6	2	48	1.6	10.9	7600	32	230000	0
7	1	70	14.1	10.9	10300	32.1	311 000	1
8	2	55	0.6	8.5	7900	27.1	211 000	0
9	1	51	1.3	13.1	31500	38.8	128000	0
10	1	50	5.1	9.5	22400	31.1	347 000	0

The dataset has 8 attributes and 200 lines. In the field of sex, 1 refers to the man and 2 refers to women. In the field of diagnosis, 0 refers to the normal and 1 refers to the diagnosis of HCC. Normal AFP value $\leq$ 7 ng/mL, normal hemoglobin value is 13-18 g/dL, normal leukocyte values are 4000-10000 (cell/uL), normal hematocrit values are 40-54 (%), and the value of normal platelets is 150000-450000 (cells/uL).

The results of the study in this paper are limited to determining whether the results of the alpha-fetoprotein blood test were diagnosed with HCC (liver cancer) or not, this study did not arrive at HCC (liver cancer) staging.

4. Results and Discussion

The following graph represents the accuracy and running time for the classification results of Hepatocellular Carcinoma (HCC) using the Possibilistic C-Means (PCM) algorithm (See **Figure 1** and **Figure2**).

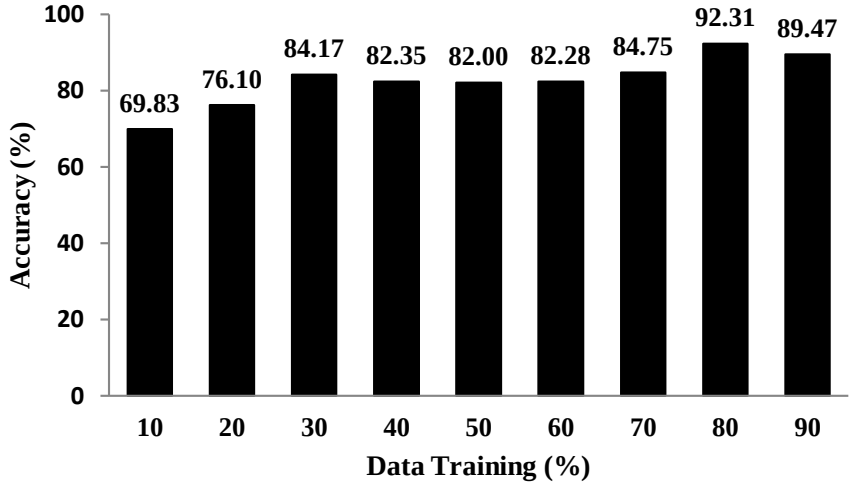

Figure 1. %Accuracy using the Possibilistic C-Means (PCM) algorithm.

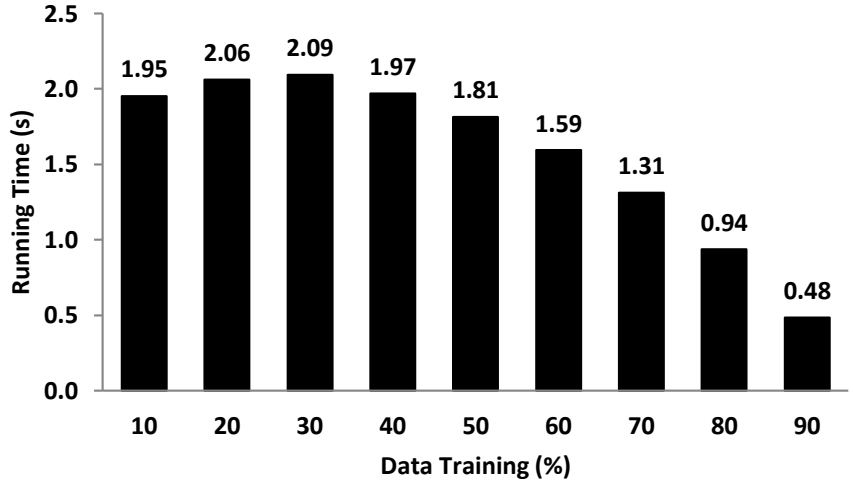

Figure 2. Running time for the Possibilistic C-Means (PCM) algorithm.

From the graph above we can see that the best accuracy is 92.31% using 80% data training. And the fastest run time is 0.48s by using 90% data training.

The following table represents other parameters that can be calculated; sensitivity, precision, and F1Score (See **Table 3**).

Table 3. Sensitivity, precision, and F1Score for the Possibilistic C-Means (PCM) algorithm.

Data Training (%)	Sensitivity	Precision	F1Score
10	54.93	63.93	59.09
20	64.81	64.81	64.81
30	90.32	59.57	71.79
40	82.76	60.00	69.57
50	90.00	52.94	66.67
60	93.33	51.85	66.67
70	100.00	55.00	70.97
80	100.00	76.92	86.96
90	100.00	66.67	80.00

From the table above we can see that the best sensitivity is 100% using 70%, 80%, and 90% data training. Then the best precision is 76.92 by using 80% data training. And the best F1Score is 86.96 by using 80% data training.

5. Conclusion

From the experiments above, we can conclude that the Positive C-Means (PCM) algorithm for Hepatocellular Carcinoma (HCC) classification has high accuracy. The best accuracy we get is 92.31%, using 80% data training, and the fastest running time is 0.48s using 90% data training. In the future, we must use a larger data set to show that the Positive C-Means (PCM) algorithm can be used as a reference to assist doctors in their final decisions, and a little modification is needed to get better accuracy.

Acknowledgment

This research was financially supported by the University of Indonesia, with PITTA B 2019 research grant scheme (ID number NKB-0688/UN2.R3.1/HKP.05.00/2019). We thank our colleagues; leaders, doctors, and laboratories at Al Islam Bandung Hospital, Bandung, West Java for providing insight and expertise that greatly helped this research.

Reference

[1] World Health Organization. *Fact Sheets: Cancer*. Accessed on February 27, 2019.https://www.who.int/news-room/fact-sheets/detail/cancer
[2] American Cancer Society. *What is Cancer ?*. Accessed on February 24, 2019.https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html
[3] American Cancer Society. *About the Liver*. Accessed on February 24, 2019.https://www.cancer.org/cancer/liver-cancer/about/what-is-liver-cancer.html
[4] LP Waller, V Deshpande, N Pyrsopoulos. *Hepatocellular Carcinoma: A Comprehensive Review*. World Journal of Hepatology. 2015.
[5] F Bravi, et al. *Coffee Reduces Risk For Hepatocellular Carcinoma: An Updated Meta-analysis*. Clinical Gastroenterology and Hepatology. 2013.
[6] A Forner, M Reig, J Bruix. *Hepatocellular Carcinoma*. The Lancet; London. 2018.
[7] G Chen, et al. *Past HBV Viral Load as Predictor of Mortality and Morbidity from HCC and the Chronic Liver Disease in A Prospective Study*. American Journal of Gastroenterology. 2006.
[8] P Dongiovanni, S Romeo, L Valenti. *Hepatocellular Carcinoma in Nonalcoholic Fatty Liver:*

Role of Environmental and Genetic Factors. World Journal of Gastroenterology. 2014.
[9] Mayo Clinic. *Diseases and Conditions: Liver Cancer*. Accessed on February 27, 2019. https://www.mayoclinic.org/diseases-conditions/hepatocellular-carcinoma/cdc-20354552
[10] American Cancer Society. *Liver Cancer Early Detection, Diagnosis, and Staging*. Accessed on February 24, 2019. https://www.cancer.org/cancer/liver-cancer/detection-diagnosis-staging.html
[11] R Krishnapuram, JM Keller. *A Possibilistic Approach to Clustering*. IEEE Transactions on Fuzzy Systems. 1993.
[12] JS Zhang, YW Leung. *Improved Possibilistic C-Means Clustering Algorithms*. IEEE Transactions on Fuzzy Systems. 2004.
[13] AVC Franca, et al. *Diagnosis, Staging and Treatment of Hepatocellular Carcinoma*. Brazilian Journal of Medical and Biological Research. 2004.
[14] AI Gomaa, MS Hashim, I Waked. *Comparing Staging Systems for Predicting Prognosis and Survival in Patients with Hepatocellular Carcinoma in Egypt*. PLoS One. 2014.
[15] B Xia, et al. *A Novel Hepatocellular Carcinoma Image Classification Method Based on Voting Ranking Random Forests*. Computational and Mathematical Methods in Medicine. 2016.
[16] AW Lestari, Z Rustam. *Kernel normed Possibilistic Function-Based Fuzzy C-Means (NKFPCM) Algorithm for High-Dimensional Breast Cancer Database Classification with Feature Selection is based on Laplacian Score*. AIP Conference Proceedings. 2017.
[17] Z Rustam, AS Talita. *Fuzzy Kernel K-Medoids Algorithm for Multiclass Multidimensional Data Classification*. Journal of Theoretical and Applied Information Technology. 2015.
[18] Z Rustam, AS Talita. *Fuzzy Kernel K-Medoids Anomaly Detection Algorithm for Problems*. International Symposium on Current Progress in Mathematics and Sciences. 2016.
[19] Z Rustam, AS Talita. *Fuzzy Kernel K-Medoids Algorithm for Anomaly Detection Problems*. AIP Publishing. 2017.
[20] Q Zhang, Z Chen. *A Distributed Weighted Possibilistic C-Means Clustering Algorithm for Incomplete Big Sensor Data*. International Journal of Distributed Sensor Networks. 2014.
[21] D Li, H Gu, L Zhang. *A Fuzzy C-Means Clustering Algorithm Based on Nearest-Neighbor Intervals for Incomplete Data*. Expert Systems with Applications. 2010.
[22] DMW Powers. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies. 2011.
[23] T Fawcett. *An Introduction to ROC Analysis*. Pattern Recognition Letters. 2006.