Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 3, May : 2023

DATA VISUALIZATION AND CLUSTER ANALYSIS OF BREAST CANCER USING WISCONSIN DATASET: AN OBSERVATION ON HIDDEN PATTERNS USING MACHINE LEARNING

Ayyakkannu Selvaraj , University Department of Information and communication technology (UDICT), MGM University, Aurangabad, Maharashtra, India, aselvaraj@mgmu.ac.in
Satheesh Kumar, Park College of Engineering and Technology, Coimbatore, Tamil Nadu, India
Doss Prakash Sundarajan, Department of Community Physiotherapy, MGM Institute of Physiotherapy, Aurangabad, Maharashtra, India

Abstract
The breast cancer has been found in women worldwide frequently and it is one of the most common cancers. The early detection is highly needed for possible cure, and hence many number of studies are ongoing in machine learning. In present study, it is aimed that to extract the hidden part from the given dataset, it can be achieved using unsupervised learning, K mean clustering is one of the most common clustering method for deriving the labeled information from unlabeled information. This study further aimed to find the best k value for clustering the dataset, and also performed the computation such as centroid distance, epoch, and iteration to cluster the data. Wisconsin dataset is concerned the parameter such as mean radius and mean texture are deployed for clustering. Since four optimized K value has been found, there has been four unlabeled information derived as hidden parts. It has been reported that whenever patient has high mean texture and mean radius, those patients have malignant type cancer.

Keywords: Clustering, Centroid, WCSS, Hidden pattern, K mean,

1. Introduction

The Breast cancer is an unavoidable disease, which makes in the breast. It can initiate in one breast or both breasts. The Breast cancer happens entirely in women, however men can get breast cancer [1]-[2]. It is noted that all breast cancer are begins in the milk ducts. Breast cancer commonly found in women. The Cancer implication starts while cancer cells begin to growing up beyond the control. [3]-[6] .The process of breast cancer would not occur in a specific day or week or in a month or even year. But the breast cancer can occurs, when there is a mutation presence [4]-[5]. The Breast cancer is a common type of cancer which can be seen in female's regardless of age group, this can be infected since various factor like lift style, genes etc. among the female.[7]-[9]. Besides Obesity is another risk of factor for developing breast cancer, dense breast tissue will lead absolutely to develop breast cancer. Due to increase health issues particularly the breast cancer has initiated to involve plenty of researcher to make further development diagnostic model. It has been observer that how to classify like breast cancer is benign or malignant with prediction of presence of cancer using conventional methods. It has been received the inference of various machine learning techniques such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes to classify the breast cancer with achieved accuracy.[6] It reported that KNN technique gives the best results rather than others learning model to identify the cancer. The SVM is an efficient method are used for recurrence/non-recurrence prediction of breast cancer, in their works SVM with multiclass provided 74% accuracy. [5]. The Breast cancer risk assessment and early diagnosis model that is accomplished with accurately establishing Breast cancer at the early stage. The data dimensionality like PCA is being used to extract features in preprocessing stage itself [10] and the features has been reduced for data processing. The multi pre-processed data were assessed for breast cancer's risk and diagnosis using SVM. The observation that was concluded that model has given good accurate when suitable

pre- processing technique deployed to pre-process the breast cancer data for feature extraction. The main objective of the present study is to deploy the K mean clustering machine learning algorithms to cluster the datasets, which is enhanced to support patient to make early diagnosis decision with help of doctor suggestion.[10]

2. Materials and Method

The Unsupervised model is the learning of computer algorithms that can improve the spontaneously by experience with data. Un Supervised learning provided the unlabeled and hidden pattern. Many researchers is used breast cancer dataset from UCI repository. Our aims to focus on certain types of clustering algorithm performance on chosen dataset. The dataset is segregated and zipped using the comment such as X = np.array(list(zip(mean texture, mean radius))).

Finding optimized k values are another important steps in clustering techniques which can be performed as follows

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1,11):
    km=KMeans(n_clusters=i, random_state=0)
    km.fit(X)
    wcss.append(km.inertia_)
plt.plot(range(1,11),wcss,color="red", marker ="8")
plt.title('Optimal K Value')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

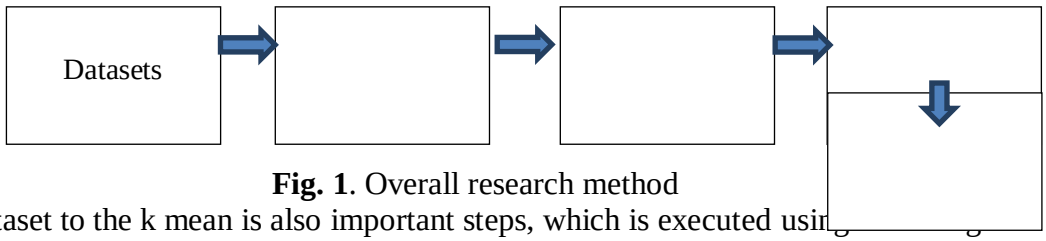

Fig. 1. Overall research method

Fitting the dataset to the k mean is also important steps, which is executed usi comments.

```
model=KMeans(n_clusters=4, random_state=0)
y_means = model.fit_predict(X)
```

Result and discussion:

The dataset can be visualized through seaborn library using python, it has been noted that when there parametric value changes with variation, seems to have breast cancer, the figure showed that when radius mean and area mean increases, it has been observed that` the cancer types premalignant (pre-cancerous) is lower than malignant. And hence it has concluded that the Malignant types is rapidly spread in nature in the body, further it can be spread to other parts and impacted others. The categorical plot has been also made Figure 2, 3

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 5, No. 3, May : 2023

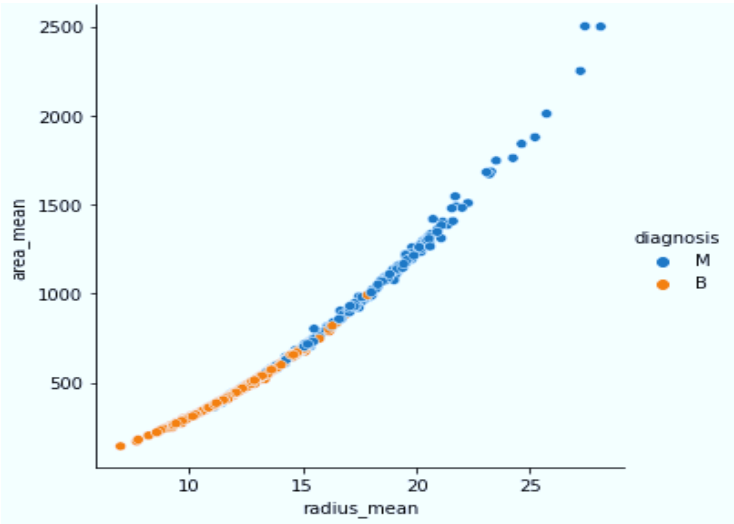

Fig. 2. Visualization of relational plot

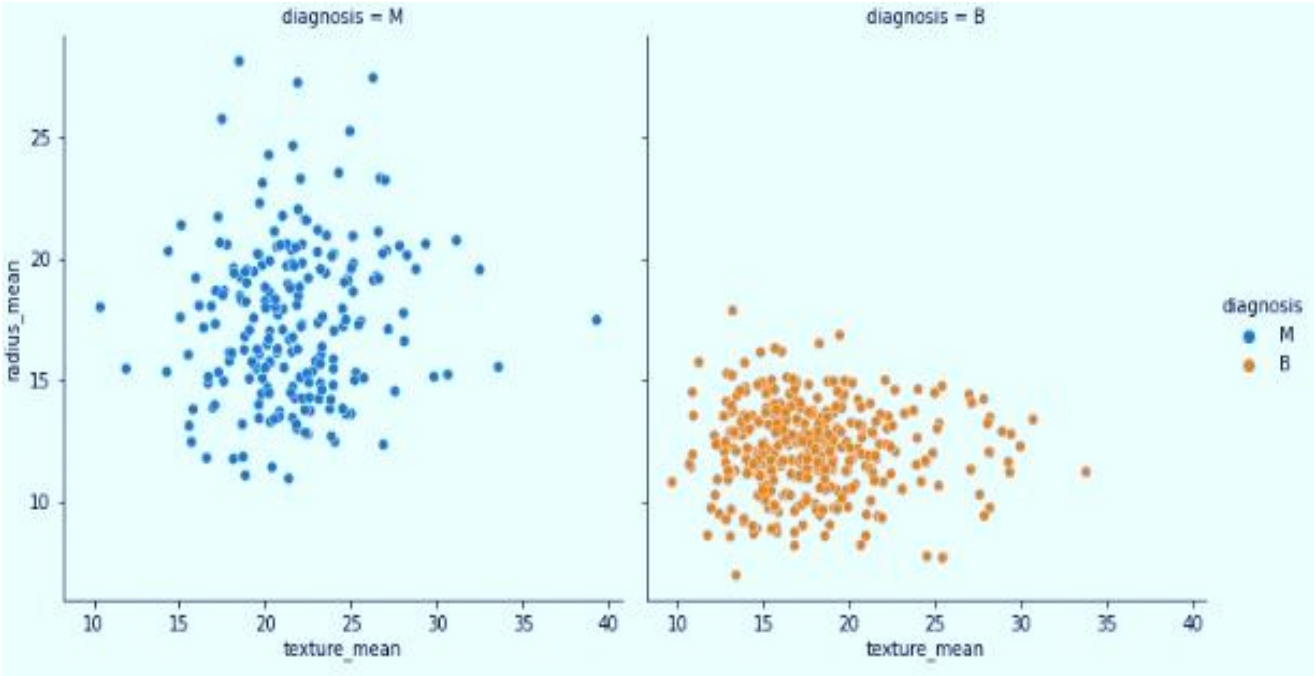

Fig. 3. Categorical plot

The result can be observed that the K mean clustering performed well on the breast cancer datasets and four hidden parts can be further extracted such as when mean texture is low and mean texture is also low, sometimes, mean texture is low but mean radius is high. Both mean texture and mean radius are high will leads to cancer (Figure 4). The hidden pattern is obtained with clusters for k=4, **Cluster** 1: Patient with medium mean texture and mean radius, **Cluster 2**: Patient with high texture and high mean radius, **Cluster** 3: Patient with low mean texture and mean radius, **Cluster** 4: Patient with high mean texture but medium mean radius.

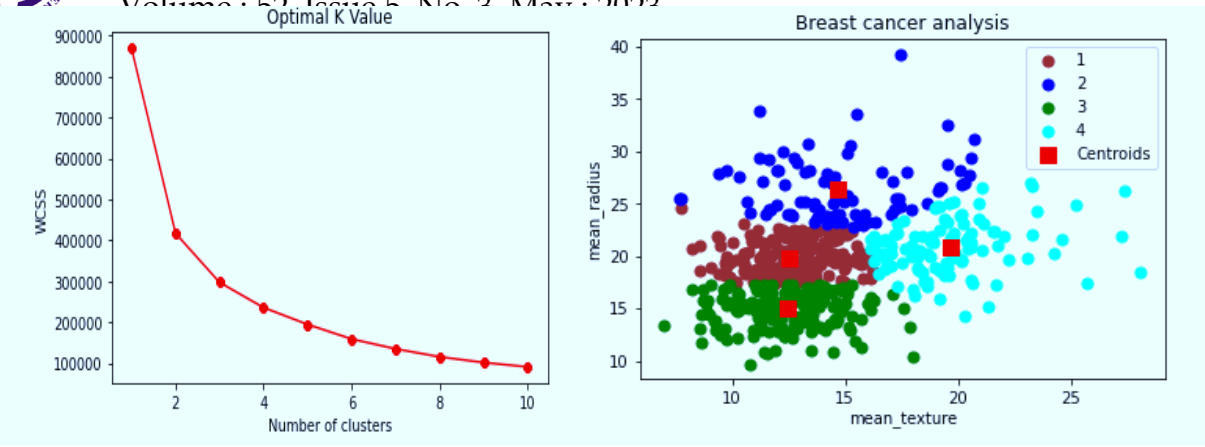

Fig. 4. Within Sum of square for optimized K value and deriving labeled information

Conclusion

In the present study, with the help of various machine learning algorithms like K mean clustering algorithms deployed to predict breast cancer with reasonable accuracy. Our results show that k mean clustering algorithms can categorize breast cancer outcomes with high accuracy and identify key characteristics even for small datasets with K = 4, it has been observed that the K mean model to be most successful clustering techniques and extracted with 4 hidden parts from the given datasets.

References

[1] Boluwaji A. Akinnuwesi, Babafemi O. Macaulay, Benjamin S. Aribisala,"Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques", ELSEVIER,pp.1-13, Oct 2020.

[2] Zexian Huang,Daqi Chen,"A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm",IEEE,Volume.10,pp.3-10,2022.

[3] Ahn, S., Woo, J. W., Lee, K., & Park, S. Y. (2020). HER2 status in breast cancer: changes in guidelines and complicating factors for interpretation. Journal of pathology and translational medicine, 54(1), 34-44

[4] Saha, M., & Chakraborty, C. (2018). Her2Net: A deep framework for semantic segmentation and classification of cell membranes and nuclei in breast cancer evaluation. IEEE Transactions on Image Processing, 27(5), 2189-2200.

[5] Jang, J.; Lee, S.; Hwang, H.J.; Baek, K. Global thresholding algorithm based on boundary selection. In Proceedings of the 2013 13th International Conference on Conference: Control, Automation and Systems (ICCAS), Gwangju, Korea, 20–23 October 2013; pp. 1–3

[6] Li, M.J.; Ng, M.K.; Cheung, Y.; Huang, J.Z. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. IEEE Trans. Knowl. Data Eng. 2008, 20, 1519–1534. [CrossRef]

[7] Kaur, P.; Singh, G.; Kaur, P. Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. Inform. Med. Unlocked 2019, 16, 100151. [CrossRef]

[8] Al-masni, M.A.; Park, J.M.; Gi, G.; Kim, T.Y.; Rivera, P. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput. Methods Programs Biomed. 2018, 157, 85–94. [CrossRef]

[9] Vaka, A.R.; Soni, B.; Reddy, K.S. Breast cancer detection by leveraging machine learning. ICT Express 2020, 6, 320–324. [CrossRef]

[10] Karabatak, M. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement 2015, 72, 32–36. [CrossRef]