International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 **www.ijisae.org** **Original Research Paper**

An Empirical Evaluation of Clustering Techniques for the Oral Cancer Prediction

S.Sivakumar [#1],T.Kamalakannan [#2]

Submitted: 05/02/2024 **Revised:** 13/03/2024 **Accepted:** 19/03/2024

Abstract-Data Mining is now widely used in healthcare applications to predict various cancers such as breast, kidney, thyroid, Colorectal, ovarian and many others. Clustering in Data Mining offers a solution for determining the prediction of Oral Cancer. This research explores K-means algorithm and introduces a new novel algorithm, the Kohonen map with K-means (Koho K-means). The experimental findings are based on 3004 oral cancer datasets, focusing on the time complexity and accuracy of the algorithms. The comparative study is then conducted with varying cluster points. The experimental results prove that Koho K-means outperforms K-means in predicting oral cancer, particularly in terms of accuracy.

Keywords – Data Mining, K-means, Koho K-means, Kohonen Map, Clustering, Oral Cancer, R and Python

1.Introduction

Three out of every four of the 3.5 billion people who suffer from oral disorders worldwide, according to the WHO Global Report 2022 [11], reside in middle-income countries. Globally, 2 billion adults have permanent dental caries, compared to 514 million children who have primary dental caries. Cancer results in excessive consumption of alcohol and tobacco products such stogies, pipes, cigarettes, cigars, betel nuts, gutka, and snuff. More than twice as many males as females suffer from oral cancer, which most often affects adults over 40 years. In this study, we have identified the most effective method for detecting the early stages of oral cancer or for treating oral cancer. Similar to other cancers, oral cancer is treated surgically to remove the abnormal growth, then with radiation therapy or chemotherapy to eradicate any residual cancer cells.

Data mining [DM] plays an important role in healthcare industries, educational analysis, credit card fraud detection, market basket analysis and many other areas. DM is the process of extracting valuable information from a vast volume of information, frequently from a group of linked data sets or data warehouse. Various data mining techniques used to mine data for different data science applications such as classification, clustering, regression, association rule mining and others. Clustering algorithm is very useful for determining the oral cancer prediction and involves the collection of information on entities that are alike to one another, placed in the identical cluster but distinct from entities in other clusters. Various clustering approaches, including K-means [10] [15] and Koho K-means are analyzed.

The remaining work are divided as follows: Section II outlines the review of numerous data mining papers for the prediction of disease. Section III provides the materials and methods of clustering techniques for oral cancer assessment. Section IV discusses the experimental findings for determining time complexity, accuracy and visualizations of various cluster points. The study's results are presented in Section V. The paper's future development is concluded in the last part.

2.Related Works

Large data sets are analyzed using data mining to generate informative results. Data mining's primary goal is to transform a massive amount of data into information that can be used to research many healthcare sectors. For the purpose of examining survival, numerous researchers have presented various clustering methods and used various valuable healthcare application domains. To achieve high clustering accuracies in their applications is the key motivation driving their research articles. We have examined several research papers that utilize the clustering algorithms to predict diseases, including Oral cancer [8], Breast cancer [9], Lung cancer [13] and others. We focus on oral cancer for the analysis of clustering algorithms.

Integrated K-means clustering has been proposed by Songul Cinaroglu[2]. The present study uses a two-stage analysis process that includes provinces and public hospitals. Phase one compares comparable provinces using the Silhouette

[#1] Research Scholar, Department of Computer Science, School of Computing Sciences,
Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamilnadu, India.
[#2]Associate Professor, HOD, Department of Information Technology,
Vels Institute of Science, Technology and Advanced Studies, Chennai, Tamilnadu, India.

cluster validity index and k-means indicators of the welfare state. After that, public hospitals in several provincial groups are assessed for effectiveness. They have collected experimental data from 81 provinces and 688 public hospitals in Turkish Statistical Institute and the Public Hospitals. Based on similarities in the welfare state characteristics, the study has discovered that there are five groups of provinces (Sil = 0.58).

Lavanya L et al [1] has examined about various classification techniques for the analysis of Oral Cancer. Their research is followed on a system of pathology and clinical staging. The dataset of the study project relies on oral leukoplakia. The data is transformed to make it more uniform, and the correlation coefficient is used to extract the characteristics. Decision tree and random forest are used to categories the retrieved characteristics, and they are contrasted with other well-liked classification techniques such KNN, MLP, Logistic Regression and SVM. In the research, predictions are made for various stages of oral cancer and the accuracy of different classification algorithms is evaluated through cross-validation. Decision tree (83.703%) and random forest (82.553%), the two algorithms, generate results with higher accuracy.

Rui Máximo Esteves et al [4] have proposed Competitive K-means. They have introduced a new technique for parallelizing K-means++ which increases the accuracy of clustering for a large volume of dataset. Additionally, it speeds up experimentation. Their Competitive K-means utilizes Hadoop and MapReduce and is very scalable. They discovered that running the algorithm on a Hadoop cluster of 15 generated speedups of 76±C times when compared to the proposed technique.

Self-organized maps as alternatives to K-Means clustering is a proposal made by Fernando Bacao et al [5]. They have taken the Lisbon Metropolitan Area count district (ED) dataset, which totals 3968 and is described using 65 variables from the 2001 Portuguese census. Six clusters within this dataset should be considered, according to a scoping study using large OBMs and U matrices. They take a batch k-means algorithm on this data to pinpoint the precise locations and make up of these 6 clusters, and compared the outcomes with those of a 6x1 SOM. We have run the experiment 100 times with random initiations in both situations. With the k averages, the square error is 3543 with a minimum of 3528, while with the MOS, the square error is 3533 6 with a minimum of 3529. These results indicate that while the optimum clustering obtained using each method is essentially the same, SOM performs better overall and has less fluctuation in its results than k-means.

R. Prabhakaran et al [7] has proposed about classification techniques for their research. This study discusses the various processes carried out on the input photos to categories them as normal or aberrant. Through morphologically based segmentation, the tumors are divided into groups, from which features are derived. They suggest various classification methods names Naïve Bayes, Convolutional Neural Network and Support Vector Machine and those algorithms have produced accurate results. Despite being a good and strong tool, SVM and Naive Bayes are still more traditional approaches than CNN. Additionally, which have used for classification is the CNN classifier 96.15% accuracy are attained for feature extraction.

kanksha Kapoor et al [6] has compared various clustering techniques for their study. Data sorting using the Fuzzy C-means, K-means++, and K-means approaches has been demonstrated as a preliminary technique. The experimental results shows that when the number of iterations decreases, the cluster performance rapidly decreases as the quantity of data points increases. In addition, sending the sorted data through all three algorithms have accelerated the process greatly. Sorting the data points reduces the variability of the cluster center, which impacts the number of iterations and time complexity.

K. Lalithamani et al [3] suggested about Enhanced Multi-Layer Perceptron in data mining. In view of medical datasets, this research develops a machine learning approach for diagnosing oral cancer. There are many techniques used to examine the early stages of oral cancer and provide basic treatments for it. The Enhanced Multi-Layer Perceptron has 92% in this method. Apriori Algorithm has 81%, SVM has 75%. According to the research, the combination of the Enhanced Multi-Layer Perceptron algorithm yields the greatest results when it comes to detecting oral cancer at an early stage.

Self Organising Map(SOM) Hybrid and K-mean techniques for breast cancer prediction have been discussed by Haoquan Lin et al [9]. According to experimental findings, the SOM Hybrid and K-means approaches each inherit and expand upon the best qualities of the K-means and SOM algorithms. The hybrid technique is more accurate than the k-means technique, and it is not only more accurate than the regular SOM hybrid algorithm, but it also runs quicker. Although the running time and accuracy of the SOM hybrid algorithm have increased, it is still slower than the k-means method.

3.Materials and Methods

We have conducted experiments with the proposed Koho K-means [KKm] and existing K-means [Km] algorithms for taking an initiative action to avoid psychoactive drugs and predicting oral cancer. The remaining analysis takes into account a variety of inputs and cluster sizes. The spacing between data points and their midpoints is maintained for cluster formations. Several colors are used to differentiate the cluster group.

A.Data Set

We collected data on 3004 people from many hospitals, labs, and other sources for experimental purposes. We used 19 distinct oral cancer data set attributes [14][16]. These attributes include age, case, gender, site, stage, TNM, histology, alcohol, smoking, dead/alive, survival time, cause_of_death, Recurrence, Disease_Free Survival(months), FAL Score, No loci AI, No loci informative, p53 IHC and Rb IHC . We analyzed both the KKm and Km algorithms, evaluating their performance in terms of accuracy and overall performance.

B.Algorithms

A comprehensive explanation of each algorithm is provided below.

i)K-means clustering algorithm [Km]

The Km algorithm is a systematic procedure that divides the dataset into k distinct, non-overlapping subgroups, each of which is represented by single data points. It allots data points to a cluster and identifies the centroid for all cluster's minimal point. Depending on their outcomes, these centroids are formed in various locations. The algorithm is presented with a step-by-step representation in the following phases.

1.Place P points inside the space represented by the clustering objects. These are the earliest group centers
2.Assign each item to the group whose centroid is nearest.
3.After assigning all objects, compute the coordinates of the k centroids.
4.Repeat steps 2 and 3 until there is no longer any movement in the centroids.

ii)Koho K-means algorithm [KKm]

Koho K-Mean is a two steps algorithm. In first step, Self-Organizing Map (Kohonen Map) is responsible to form a large set of prototypes which are then joined to create the actual clusters in the following step. step, performs better and speeds up computation compared to doing direct data clustering. Each data vector of the original data sets are all members of the same cluster as their closest prototypes. The algorithm is presented with a step-by-step representation in the following phases.

1. Initialization of each node's weights with a random number between 0 and 1
2. Choosing a random input vector from training dataset.
3. Calculating the Best Matching Unit (BMU). Each node is examined to find the one which its weights are most similar to the input vector.
4. Calculating the size of the neighborhood around the BMU.
5. Modification of nodes' weights of the BMU and neighboring nodes, so that their weight gets more similar to the weight of input vector. The weight of every node within the neighborhood is adjusted, having greater change for neighbors closer to the BMU.
6. Obtain the output, which includes the number of clusters K and the cluster center Z ((Z1, Z2,...,Zk)). In the second step of the K-means algorithm, the results from the first stage, specifically the number of clusters K and the cluster center point Z, are utilized as the starting input values.
7. The result of step 7 serves as the starting input to the K-means algorithm for iterative computation until convergence.
8. Display the clustering results of the Koho K-means algorithm.

4.Experimental Results

In the evaluation of clustering algorithms in data mining, researchers make use of tools such as Python, R Programming, and Weka. We evaluate the time complexity, centroid point of various clusters and accuracy of both the Km and KKm algorithms through the use of R programming and Python.

We utilized the dataset to predict oral cancer in individuals aged 20 to 60 with varying smoking volumes ranging from 1 to 32. Employing two algorithms, we partitioned the dataset into multiple clusters based on the user-specified value of k. For each cluster number, we present the means of instances within the cluster, along with the count of instances and the corresponding percentage representation based on the total instances. Table 6.1 represents the centroids of 4 Cluster and the respective number of data points associated with each centroid.

Table 6.1: 4 cluster data points

Cluster No	K-means	Koho K-means
Cluster 1	854	935
Cluster 2	896	778
Cluster 3	551	471
Cluster 4	702	819

Figure 6.1 displays the centroids of 4 Cluster and the corresponding number of data points linked to each centroid.

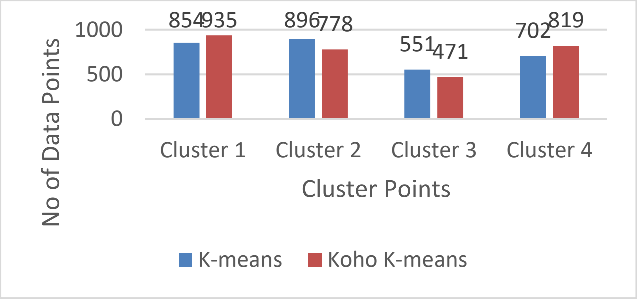

Figure 6.1: Data points of 4 Cluster

Table 6.2 represents the centroids of 6 Cluster and the respective number of data points associated with each centroid.

Table 6.2: 6 cluster data points

Cluster No	K-means	Koho K-means
Cluster 1	636	759
Cluster 2	421	434
Cluster 3	347	554
Cluster 4	770	495
Cluster 5	418	566
Cluster 6	411	195

Figure 6.2 displays the centroids of Cluster 6 and the corresponding number of data points linked to each centroid.

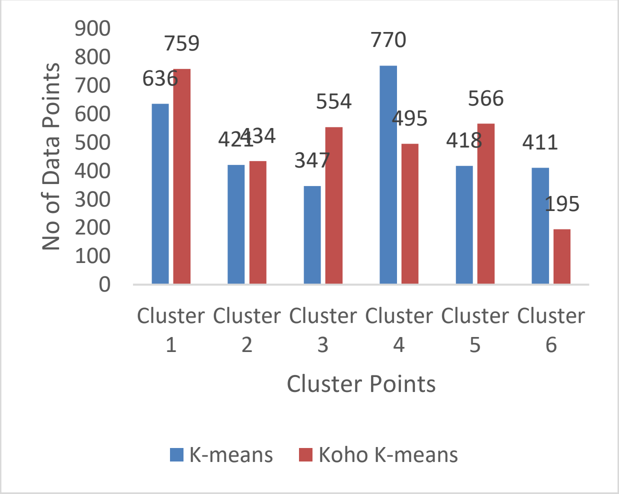

Figure 6.2: Data points of 6 Cluster

Table 6.3 represents the centroids of Cluster 8 and the respective number of data points associated with each centroid.

Table 6.3: 8 cluster data points

Cluster No	K-means	Koho K-means
Cluster 1	251	325
Cluster 2	685	620
Cluster 3	379	466
Cluster 4	486	160
Cluster 5	534	270
Cluster 6	205	379
Cluster 7	196	347
Cluster 8	267	436

Figure 6.3 displays the centroids of Cluster 8 and the corresponding number of data points linked to each centroid.

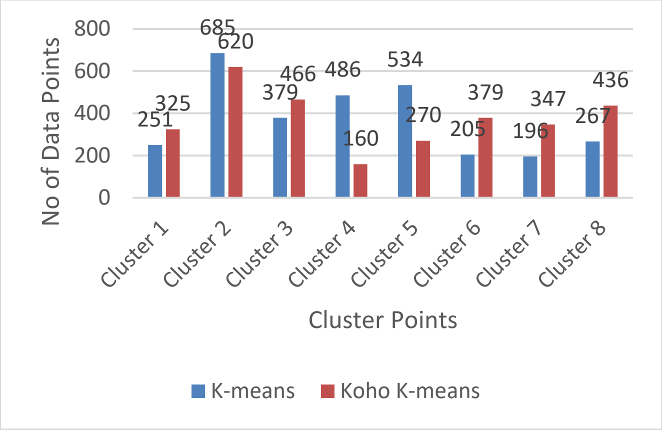

Figure 6.3: Data points of 8 Cluster

The time complexity of above two algorithms is computed with reference to the dataset used for oral cancer analysis. Analyzing datasets that focus on age ranges between 20 and 60 and smoking ranges between 1 and 32, the K-means algorithm exhibits time complexities for cluster 4, cluster 6 and cluster 8 are 6134, 5844 and 7962 milliseconds, the Koho k-means algorithm exhibits time complexities for cluster 4, cluster 6 and cluster 8 are 7321, 8765 and 9722 milliseconds.

The comparative analysis confirms that the K-means algorithm displays a lower time complexity in contrast to the Koho K-means algorithm. Table 6.4 shows the time complexities of the algorithms.

Table 6.4: Time complexities of the algorithms

Algorithms	Runtime (in Milliseconds)		
	Cluster Points		
	4	6	8
K-means	6134	5844	7962
Koho K-means	7321	8765	9722

Figure 6.4 displays the time complexities of 4 clusters, 6 clusters and 8 clusters of the algorithms.

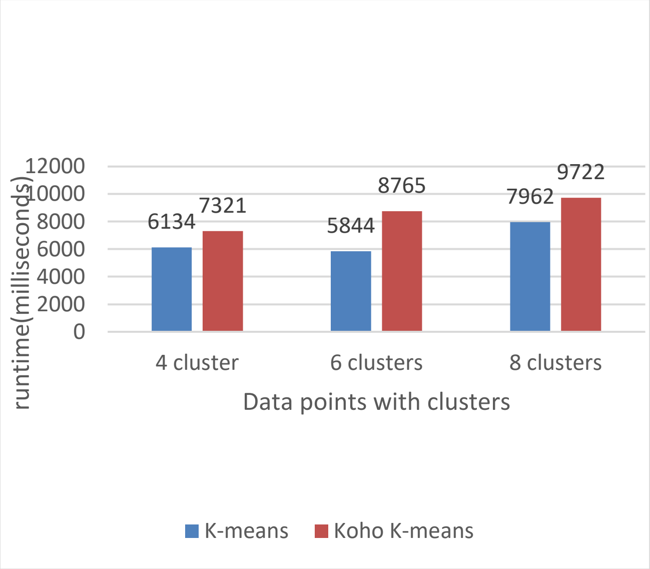

Figure 6.4: Time complexities of the algorithms

Figure 6.5 presents the outcomes of accuracy tests conducted on both the K-means and the newly proposed Koho K-means technology for clusters 4, 6, and 8. The proposed algorithm outperforms existing algorithms, notably K-Means, with higher accuracy rates: 90.8%, 90.75%, and 91.1%, compared to K-means' rates of 85.84%, 85.27%, and 85.42%. The results indicate the superiority of the proposed algorithm over the existing algorithm.

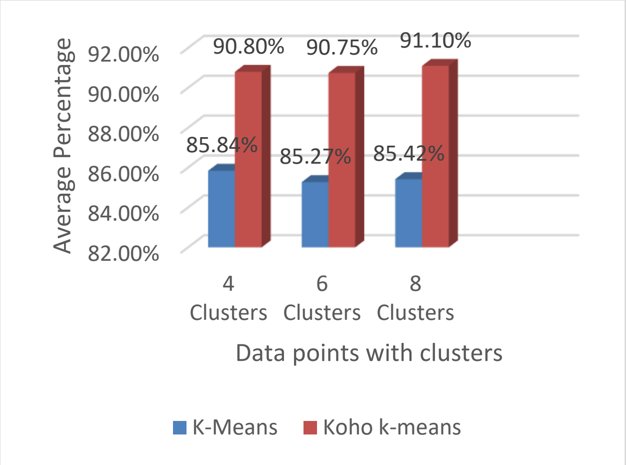

Figure 6.5: Accuracy of K-means and Koho K-means

5. Conclusion

We conducted a comparative analysis of accuracy and time complexity for oral cancer prediction between K-means and Koho K-means. The dataset was employed to assess the time complexity and accuracy of the algorithms. Analysis of testing data reveals that the Koho K-means algorithm exhibits high accuracy but with a longer running time, while the K-means algorithm shows lower accuracy but with a shorter running time. Overall, considering both accuracy and processing speed, the Koho K-means algorithm model performs better than K-means and is more suitable for predicting oral cancer. These tests provide medical experts with reliable tools for accurate oral cancer prediction.

6. Future enhancements

The objective of future development is to maintain the accuracy of Koho K-means algorithm while reducing the its time complexity. Additionally, more data must be utilized to determine the algorithm's high accuracy.

References

[1] Lavanya L and Dr. Chandra J, Oral Cancer Analysis Using Machine Learning Techniques, International Journal of Engineering Research and Technology, Volume 12, Number 5 (2019), pp.596-601.

[2] Songul Cinaroglu, Integrated k-means clustering with data envelopment analysis of public hospital efficiency, Health Care Management Science, Springer, pp. 325-338, 2020.

[3] K. Lalithamani, A. Punitha, A Machine Learning Approach for Oral Cancer Detection Using Enhanced Multi-Layer Perceptron, International Journal of Innovative Research in Applied Sciences and Engineering, Volume 2, Issue 8, pp. 319-331, February 2019.

[4] Rui Máximo Esteves, Thomas Hacker and Chunming Rong, Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets, IEEE Explore, 2014. DOI: 10.1109/CloudCom.2013.89

[5].Fernando Bação, Victor Lobo1, and Marco Painho, Self-organizing Maps as Substitutes for K-Means Clustering ,Springer, LNTCS,Volume 3516, 2005.

[6] Akanksha Kapoor and Abhishek Singhal, A Comparative Study of K-Means, K-Means++ and Fuzzy C- Means Clustering Algorithms, IEEE Explore,2017. DOI:10.1109/CIACT.2017.7977272

[7]] R. Prabhakaran, J. Mohana, Detection of Oral Cancer Using Machine Learning Classification Methods, International Journal of Electrical Engineering and Technology, Volume 11(3), 2020, pp. 384-393.

[8].Shilpa Harnale , Dhananjay Maktedar, Oral Cancer Detection: Hybrid Method of KFCM Clustering, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume 8 ,Issue 5, January 2020.

[9] Noor Kadhim Ayoob, Breast Cancer Diagnosis Using K-means Methodology, Journal of Babylon University, Pure and Applied Sciences, Number (1), Volume (26), 2018.

[10] Velmurugan T, Efficiency of K-means and K-medoids Algorithms for Clustering Arbitrary Data Points, International Journal of Computer Technology & Technology, Volume 3(5), pp. 1758-1764, 2012.

[11].World Health Organization:https://www.who.int/team/noncommunicable-diseases/global-status-report-on-oral-health-2022.

[12]The Indian Council of Medical Research: https://main.icmr.nic.in/

[13] Alka Kumari and Megha Kamble, Improved Clustering Methodology for Lung Cancer Disease Prediction, International Journal of LNCT, Volume 4, Issue 16, February 2020.

[14] Fatihah Mohd, Zainab Abu Bakar, Noor Maizura Mohamad Noor, Zainul Ahmad Rajion, Data preparation for pre-processing on oral cancer dataset, IEEE Explore 2013. DOI: 10.1109/ICCAS.2013.6703916.

[15] S.Sivakumar and T.Kamalakannan, Performance based Analysis of K-Medoids and K-Means Algorithms for the Diagnosis and Prediction of Oral Cancer, Computational Intelligence for Clinical Diagnosis, Springer, pp. 215-226, 2023.

[16] Arushi Tetarbe , Tanupriya Choudhury ,Teoh Teik Toe and Seema Rawat, Oral cancer detection using data mining tool ,IEEE Explore, pp. 35-39, 2017. ISBN:978-1-5386-1145-6