International Journal on Advanced Science Engineering Information Technology

Vol.8 (2018) No. 1
ISSN: 2088-5334

Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data

Ashutosh Kumar Dubey[#], Umesh Gupta[#], Sonal Jain[#]

[#] *Institute of Engineering and Technology, JK Lakshmipat University,*
Near Mahindra SEZ, Ajmer Road, Jaipur - 302 026, (Rajasthan) India
E-mail: ashutoshdubey123@gmail.com

Abstract*— Breast cancer is one of the most common forms of cancer having a worldwide prevalence. Continuous research is going on for detecting breast cancer in its early stage as the possibility of cure is very high in the early stage. The two main objectives of this work were: firstly, to compare the performance of k-means and fuzzy c-means (FCM) clustering algorithms; and secondly, to make an attempt to carefully consider and examine, from multiple points of view, the combination of different computational measures for k-means and FCM algorithms for a potential to achieve better clustering accuracy. K-means and FCM algorithms have been considered to understand the impact of clustering on the breast cancer data. The execution of k-means algorithm is based on centroid, distance, split method, threshold, epoch, attributes, and number of iterations; while FCM is executed on the basis of fuzziness value and termination condition. The breast cancer Wisconsin (BCW) dataset was used for the experimentation and the comparison. The combination of variance and same centroid offers better outcome in terms of k-means algorithm. The highest and lowest clustering accuracies are (94.7%, 77.1%) and (94.4%, 88.5%) for foggy and random centroid, respectively. The overall average positive prediction accuracy obtained by this approach is approximately 92%. In case of FCM, the highest and lowest clustering accuracies are (97.2%, 91.1%), (97.2%, 90.9%), (97.8%, 90.4%), and (97.1%, 90.2%) for different combination of fuzziness and termination criteria. The average highest and lowest clustering accuracies are (95.7%, 94.7%), (95.9%, 93.6%), (95.3%, 94.2%), and (95.6%, 93.7%) for the same combination in the case of FCM. K-means algorithm was more prominent and consistent in terms of computation time as FCM required more time to carry out several fuzzy calculations and iterations. The findings of this work provide an incisive and extensive understanding of the computational parameters used with k-means and FCM algorithms. The computational results indicate that FCM algorithm was found to be prominent and consistent than k-means algorithm when executed with different iterations, fuzziness values, and termination criteria. It is more potentially capable in clustering BCW dataset as the clustering accuracy is more important than time.

***Keywords*— Breast cancer; breast cancer Wisconsin dataset; k-means; fuzzy c-means.**

I. INTRODUCTION

Breast cancer is the most frequently diagnosed cancer in women [1]. There are 458,000 deaths per year from breast cancer worldwide, making it the most common cause of female cancer with a high mortality in both the developed and developing countries [1, 2].

An early diagnosis of the breast cancer can be helpful as the chances of a complete cure are high [3, 4]. Its symptoms may vary according to the conditions, so the features identification is important. In this regard, the pattern detection is very important so that even hidden patterns can be identified correctly. Data mining techniques are capable in identifying the hidden patterns [4]; and can efficiently be used in classification, estimation, prediction, association rules, clustering, and visualization [5].

Prediction, classification, and estimation are included in the supervised learning category. In these techniques, clustering is important for data grouping as it is capable to cluster the data based on the property or symptom of the disease. K-means or hard c-means and fuzzy c-means are the mostly used clustering algorithms. The main benefit of k-means algorithm is that if the k (number of cluster) is small then the achieved computational speed is high even for large variables. The use of k-means algorithm is increasing day by day in the field of medical research because of its better clustering capabilities. It is basically a partitioning method applied to analyze data and to treat the observations of the data as objects based on locations and distance between the various input data points [6]. Mary et al. [7] also used k-means algorithm for cluster point refinement and used ant colony optimization (ACO) for cluster quality improvement. Wang et al. [8] formulated a clustering method named