Int J CARS
DOI 10.1007/s11548-016-1437-9

ORIGINAL ARTICLE

Analysis of k-means clustering approach on the breast cancer Wisconsin dataset

Ashutosh Kumar Dubey[1] · **Umesh Gupta**[1] · **Sonal Jain**[1]

Received: 15 February 2016 / Accepted: 27 May 2016

Abstract
Purpose Breast cancer is one of the most common cancers found worldwide and most frequently found in women. An early detection of breast cancer provides the possibility of its cure; therefore, a large number of studies are currently going on to identify methods that can detect breast cancer in its early stages. This study was aimed to find the effects of k-means clustering algorithm with different computation measures like centroid, distance, split method, epoch, attribute, and iteration and to carefully consider and identify the combination of measures that has potential of highly accurate clustering accuracy.
Methods K-means algorithm was used to evaluate the impact of clustering using centroid initialization, distance measures, and split methods. The experiments were performed using breast cancer Wisconsin (BCW) diagnostic dataset. Foggy and random centroids were used for the centroid initialization. In foggy centroid, based on random values, the first centroid was calculated. For random centroid, the initial centroid was considered as (0, 0).
Results The results were obtained by employing k-means algorithm and are discussed with different cases considering variable parameters. The calculations were based on the centroid (foggy/random), distance (Euclidean/Manhattan/Pearson), split (simple/variance), threshold (constant epoch/same centroid), attribute (2–9), and iteration (4–10). Approximately, 92 % average positive prediction accuracy was obtained with this approach. Better results were found for the same centroid and the highest variance. The results achieved using Euclidean and Manhattan were better than the Pearson correlation.
Conclusions The findings of this work provided extensive understanding of the computational parameters that can be used with k-means. The results indicated that k-means has a potential to classify BCW dataset.

Keywords Breast cancer · Breast cancer Wisconsin (BCW) diagnostic dataset · K-means · Foggy and random centroid

Introduction

Breast cancer, the second most common cancer across the world after lung cancer, is by far the most frequent cause of cancer death in women [1,2]. If it is diagnosed in early stages, the possibilities of survival are higher [3]. Since its symptoms vary from patient-to-patient, it is essential to characterize distinctive features of different patients and design a patient-specific treatment. The detection of the pattern of symptoms using data mining is a very important technique to correctly understand hidden patterns. The pertinent patterns extraction from the huge database is possible because of data mining techniques [4]. According to Jain et al. [5], data mining can be used for classification, estimation, prediction, association rules, clustering, and visualization activities. Of these activities, prediction, classification, and estimation come in supervised learning categories that prepare the model based on the available data representing one or more attributes. In these techniques, clustering is an important activity that enables grouping of data based on the nature or a symptom of the disease. So it can be applied in the primary stage for data pruning. K-means algorithm is one of the simple and important clustering algorithms. Classification

✉ Ashutosh Kumar Dubey
ashutoshdubey123@gmail.com

[1] JK Lakshmipat University, Near Mahindra SEZ, P.O. Mahapura Ajmer Road, Jaipur, Rajasthan 302 026, India

Published online: 16 June 2016