IPRC/16/66

Optimization of SpdK-means Algorithm

R.P.T.H. Gunasekara[1*], M. C.Wijegunasekara[2], N. G. J. Dias[2]

This study was carried out to enhance the performance of the *k*-mean data-mining algorithm by using parallel programming methodologies. As a result, the Speedup *k*-means (SpdK-means) algorithm which is an extension of *k-means* algorithm was implemented to reduce the cluster building time. Although SpdK-means speed up the cluster building process, the main drawback was that the cumulative cluster density of the created clusters by the SpdK-means algorithm was different from the initial population. This means some elements (data points) were missed out in the clustering process which reduces the cluster quality. The aim of this paper is to discuss how the drawback was identified and how the SpdK-means algorithm was optimized to overcome the identified drawback.

The SpdK-means clustering algorithm was applied to three datasets which was gathered from a Ceylon Electricity Board Dataset by changing the number of clusters *k*. For k=2, 3, 4 did not give any significant difference between the cumulative cluster density and the initial dataset. When the number of clusters were more than 4 (i.e., when k>=5), there was a significant difference on cluster densities. The density of each cluster was recorded and it was identified that the cumulative density of all clusters was different from the initial population. It was identified that about 1% of elements from total population were missing after clusters were formed.

To overcome this identified drawback the SpdK-mean clustering algorithm was studied carefully and it was identified that there are elements which had equal distances from several cluster centroids were missed out in intermediate iterations. When an element had an equal distance to two or more centroids the SpdK-means algorithm was unable to identify to which cluster that the element should belong and as a result the element is not included in any cluster. If such element was included into all the clusters that had an equal distance and if this process is repeated to all such elements the cumulative cluster density will be highly increased from the initial population.

Therefore, the SpdK-means was optimized by selecting one of the cluster centroids which had equal distance to one element. After many studies of selection methods and their outcomes, it was able to modify the SpdK-means algorithm to find suitable cluster to an equal distance element. Since, an element can belong to any cluster it is not possible give any priority to select a belonging cluster. As all centroids had equal distances from the elements, the algorithm will select one of the centroid from all equal centroids randomly.

The developed optimized SpdK-means algorithm successfully solved the identified problem by identifying missing elements and including them in to the correct clusters. By analyzing the iterations when applied to the datasets, the number of iterations was reduced by 20% than the former SpdK-means algorithm. After applying optimized SpdK-means algorithm to above mentioned datasets, it was found that it reduces the cluster building time by 10% to 12% than the SpdK-means algorithm. Therefore, the cluster building time was further reduced than the former SpdK-means algorithm.

[1] Wayamba University of Sri Lanka. *hansigunasekara123@gmail.com
[2] University of Kelaniya, Sri Lanka.