ISSN: 1391-8796
Proceedings of
2nd Ruhuna International Science & Technology Conference
University of Ruhuna, Matara, Sri Lanka
January 22-23, 2015

A study on clustering algorithms in data mining using Weka tool

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

[1]*Department of Computing and Information Systems, Wayamba University of Sri Lanka.*
[2]*Department of Statistics & Computer Science, University of Kelaniya*

This study is based on clustering data mining algorithms by using WEKA machine learning software. This paper discusses about four clustering algorithms: *k*-means, Expectation Maximization(EM) ,Density Based and Hierarchical clustering algorithm, and study the performance of these clustering algorithms based on the cluster building time of each algorithm and the quality of built clusters. The experiment is done on five datasets using WEKA interface. In this experiment, the selected four clustering algorithms are used for five datasets to create clusters. From the results obtained in the experiment, it was concluded that there are both advantages and disadvantages among these clustering algorithms. The *k*-mean significantly reflected that it is the best performing algorithm for large datasets and cluster building time taken was significantly low. Density based clustering algorithm was not suitable for data with high variance in density. Hierarchical Clustering algorithm did not support for large datasets. However Hierarchical clustering algorithm was more sensitive for noisy or outlier data. EM clustering algorithm gave log likelihood values of the clusters to ensure more reliable clusters. EM algorithm is an extension of *k*-mean which satisfies more iterations. Although this is a complex algorithm, it can be applied to parallelization to obtain best performances using cross validation. According to this study, it was identified that to choose the best clustering algorithm it is necessary to study size of the dataset, density of the dataset and its distribution. This study is continued for several clustering algorithms to increase the performance by using parallel programming methodologies.

*hansigunasekara@yahoo.com