Gunasekara, R.P.T.H.
Dias, N.G.J.
Wijegunasekara, M.C.
PAPER

Performance of *k*-mean data mining algorithm with the use of WEKA-parallel

R.P.T.H. Gunasekara, Department of Computing & Information Systems, Wayamba University, N.G.J. Dias & M.C. Wijegunasekara, Department of Statistics & Computer Science, University of Kelaniya

This study is based on enhancing the performance of the *k*-mean data mining algorithm by using parallel programming methodologies. To identify the performance of parallelizing, first a study was done on *k*-mean algorithm using WEKA in a stand-alone machine and then compared with the performance of *k*-mean with WEKA-parallel.

Data mining is a process to discover if data exhibit similar patterns from the database/dataset in the different areas like finance, retail industry, science, statistics, medical sciences, artificial intelligence, neuro science etc. To discover patterns from large data sets, clustering algorithms such as *k*-mean, *k* -medoid and, balance iterative reducing and clustering using hierarchies (BIRCH) are used. In data mining, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* (where *k* is the number of selected groups) clusters in which each observation belongs to the cluster with the nearest mean. The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid (Center of Mass of the cluster).

As the data sets are increasing exponentially, high performance technologies are needed to analyze and to recognize patterns of those data. The applications or the algorithms that are used for these processes have to invoke data records several times iteratively. Therefore, this process is very time consuming and consumes more device memory on a very large scale. During the study of enhancing the performance of data mining algorithms, it was identified that the data mining algorithms that were developed for the parallel processing were based on the distributed, cluster or grid computing environments. Nowadays, the algorithms are required to implement the multi-core processor to utilize the full computation power of the processors.

The widely used machine learning and data mining software, namely WEKA was first chosen to analyze clusters and identify the performance of *k* -mean algorithm. *k* -mean clustering algorithm was applied to an electricity consumption dataset to generate *k* clusters. As a result, the dataset was partitioned into *k* clusters along with their mean values and the time taken to build clusters was also recorded. (The dataset consists of 30000 entries and it was collected from the Ceylon Electricity Board).

Secondly to reduce the time consumed, we selected parallel environment using WEKA-parallel (Machine Learning software). This is a new option of WEKA used for multi-core programming methodology that can be used to connect several servers and client machines. Here, threads are passed among machines to fulfill this task. The WEKA parallel was installed and established for some distributed server machines with one client machine. The same electricity consumption dataset was used with *k* -mean in WEKA-parallel. The speed of building clusters was increased when the parallel software was used. But the mean values of the clusters are not exact with the previously obtained clusters. By visualizing both sets of clusters it was identified that some

border elements of the first set of clusters have jumped to other clusters. The mean values of clusters are changed because of those jumped elements.

The experiment was done on a single core i3, 3.3 GHz machine with Linux operating system to find the execution time taken to create *k* number of clusters using WEKA for several different datasets. The same experiment was repeated on a cluster of machines with similar specifications to compute the execution time taken to create *k* number of clusters in a parallel environment using WEKA-parallel by varying the number of machines in the cluster. According to the results, WEKA-parallel significantly improves the speed of *k*-mean clustering. The results of the experiment for a dataset on the consumption of electricity consumers in the North Western Province are shown in Table 1.

Algorithm & Software	**Number of connected Machines**	**Number of clusters (*k* value)**	**Execution time**
k-Mean (WEKA)	1	3	1.18 seconds
k -Mean (WEKA)	1	4	1.46 seconds
k -Mean (WEKA-parallel)	2	3	0.69 seconds
k -Mean (WEKA-parallel)	2	4	0.78 seconds
k -Mean (WEKA-parallel)	3	3	0.56 seconds
k -Mean (WEKA-parallel)	3	4	0.64 seconds

Table 1: Experiment results of Electricity consumption data

This study shows that the use of WEKA-parallel and parallel programming methodologies significantly improve the performance of the *k*-mean data mining algorithm for building clusters.