IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 6, December, 2020, pp. 763-769
Journal homepage: http://iieta.org/journals/ria

Microarray Breast Cancer Data Clustering Using Map Reduce Based K-Means Algorithm

Hymavathi Thottathyl*, Kanadam Karteeka Pavan, Rajeev Priyatam Panchadula

Department of Computer Applications, R.V.R.& J.C.College of Engineering, Chowdavaram, Guntur 522019, A.P., India

Corresponding Author Email: hyma@rvrjc.ac.in

https://doi.org/10.18280/ria.340610

Received: 8 August 2020
Accepted: 11 December 2020

Keywords:
microarray data, clustering, unsupervised learning, unlabelled data, gene expression

ABSTRACT

Breast cancer is one of the world's most advanced and most common cancers occurring in women. An early diagnosis of breast cancer offers treatment for it; therefore, several experiments are in development establishing approaches for the early detection of breast cancer. The great increase in research in the last decade in microarray data processing is a potent tool of diagnosing diseases. Based on genomic knowledge, micro-arrays have changed the way clinical pathology recognizes, identifies, and classifies the diseases of humans, particularly those of cancer. In this article, we examined microarray data for breast cancer with the k-means clustering algorithm, but it was hard to scale and process a large number of micro-array data alone. To this end, we use a chart to minimize the paradigm for evaluating microarray data on breast cancer. Moreover, the efficiency of the parallel k-means model is measured with the operating period, the scaling, and all runtime of the model.

1. INTRODUCTION

The leading cause of death in women worldwide was Breast cancer [1, 2], the second most common cancer across the world after lung cancer. The odds of recovery are better when diagnosed in the early stages [3]. As the signs vary from patient to patient, the distinguishing characteristics of various patients are crucial to characterize and patient-specific care is planned. The patient's genomic data is well-suited to extract these characters. The phenomenal increase in science in microarray data processing in the last decade is a valuable method for diagnosing diseases. Microarrays focused on genomic knowledge, changed the way clinical pathology recognizes, explains, and categorizes human diseases, particularly cancer. If detected early and correctly, cancer patients will benefit from more effective care and more resectable tumors. DNA microarrays yield large amounts of genetic data that are theoretically valuable for cancer detection and comprise of meaningless and noisy data. Data sets are degraded by the existence of obsolete, irrelevant, and distracting genomes. In the design of a diagnostic model for the disease, approaches to gene selection particularly when samples are small.

To correctly interpret secret trends, the recognition of signs by way of data extraction is a very necessary technique. Data mining techniques [4] allow it possible to remove the related trends from the vast database. Data mining can be used for classifying, forecasting, projecting, associating laws, clustering, and visualizing practices in conjunction with Devi, R.D.H. and Devi, M.I. [5]. The prediction, description, and evaluation of these events are in categories of supervised learning that prepare the model based on data describing one or more attributes accessible. The classification of data depending on the existence or the symptom of the condition is an essential task in these strategies. It may also be used for data pruning in the primary process.

One of the basic and essential clustering algorithms is the K-means algorithm. The advantages of K-Means make it famous. The most significant ones are flexibility and ease of use [2]. Besides, it has linear complexity in space and is generally fast. K-means even has a lot of inconveniences. K-mean's deterministic character is one of the major disadvantages. K-means begins with a random number of data center points. This random sorting influences the consistency of the clusters that arise. Classification can be dependent on parametric, semi-parameter, or not. The parametric method is focused on a known distribution template, a non-parametric distribution sample, and a semi-parametric distribution sample from both a known distribution and an uncertain distribution [6, 7]. K-means the clustering of k-clusters is a semi-parametric process and a simpler way to cluster k-clusters. K-means have the key benefit that if the number of clusters is tiny, it can be fast machine speed for the major element.

However, it takes a lot of time to work with large quantities of microarray data to do this, utilizing the map reduction models parallel processing technique. Map Reduce [2] is a programming model that facilitates the dissemination of big data on the commodity cluster. Micro array data is an unstructured data, in order to handle that data we use parallel processing technique of Map-reduce and K-means is used for clustering of the features of the micro array data. Figure 1 shows how the map-reduce model is work. This architecture has been commonly used for carrying out data-intensive work in texts/graphs, machine learning, and the bioinformatics industry thanks to its attractive characteristics, including scalability, simplicity, and tolerance for faults [3, 4]. Simply, each data block in Map Reduce is allocated to a working node, which can be preserved on all distributed servers (e.g., HDFS in Hadoop [5]). Input blocks are mapped to those intermediate pairs (schlüssle, value). In the next stage, known as data mixing, the intermediate couples (key, value) are passed to a variety of processors via inter-server communication links to

reduce the results to the final ones. One of the major flashpoints for improving the Map Minimize efficiency is the data shuffling process.

In this paper, we use a map-reduce based k-means clustering algorithm for clustering of microarray Brest cancer data. The methodology is better to cluster the relevant data and made clear that for the detection of breast cancer. Moreover, the model is run with a good amount of time and produces better results than existing models. The rest of the article is organized as follows section-2 gives the details of literature, section-3 presents the map-reduce-based k-means model for the analysis of breast cancer data, section-4 produces the experimental results and finally section-5 concludes the paper.

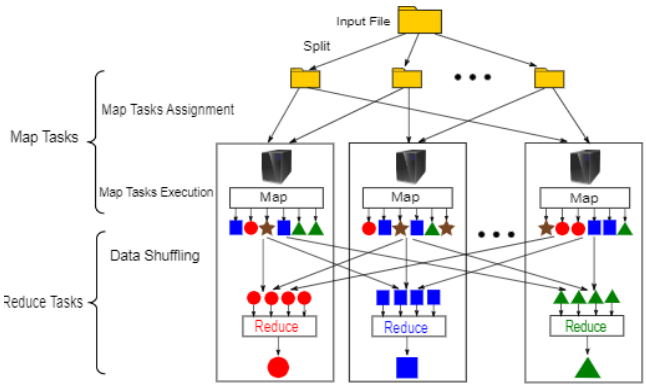

Figure 1. Map-reduce model

2. LITERATURE SURVEY

Ahmad et al. [1] uses SVM, Decision Tree (C4.5), and ANN to build and evaluate their performance based on sensitivity, precision, and accuracy to develop the predictive model. This indicates that high-precision SVM forecasts [2] test SVM and Bayesian survival forecasts for the breast cancer patient, and that 74.44Bayesians achieved a total of 67.56Genetic (GA) and Shared Knowledge (MI) for Breast Cancer Diagnostic Network, out of the Bayesian network. The outcome was obtained by 73.44 Bayesian Nets. This hybrid solution uses GA as an entry to the support vector machine (SVM) and the k-Nearest neighbor (k-NN) for choosing the ideal selection of functions. MI is used to improve reciprocal knowledge between characteristics. The experimental findings show that the method suggested for cancer forecasting is extremely successful and can be helpful for doctors. To decide the correctness of data classification in terms of the performance, accuracy, and efficacy of each algorithm, Asri et al. [3] compares various machine learning algorithms, including Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbours (k-NN) data on the Wisconsin Breast Cancer (original data) The outcome indicates SVM with the lowest error rate to be highly reliable (97.13 percent). The genetic algorithm is applied for 3- fold-over-validation, offering a clustering precision of 97.7 percent that is precise relative to current algorithms. The genetic algorithm is used for 3-fold cross-validation. Bethapudi et al. [4] based on automated breast cancer detection using machine learning techniques. In three steps, the suggested model is implemented: first, the creation of clusters using the Farthest First Clustering algorithm. Secondly, the outliers are identified using an Outliers Identification algorithm and lastly a cancer classification using a J48 classification algorithm, be it healthy or malignant. Testing on WBCD and WDBC demonstrates the reliability of the model. The numerical review reveals that the optimal possible formula for the model is 99.9%, WBCD 99.6%, and WDBC 99.6%, respectively. A hybrid diagnostic method, Elouedi et al. [6] suggested maximizing the classification of malignant cases using the Decision Tree (C4.5). Clustering and K-mean (K-mean). The subdivision of two clusters of malignant instances increases the cancer preview outcome according to the original findings. The generic classification model is introduced by Elshazly et al. [7], which relies on a rough method and judgment laws. This approach is used for evaluating three classification strategies (Equal binning and entropy and Boolean reasoning) in a mixture of 3 separate classifications (Decision Tree (DT), K Next neighbor (KNN), and Naive Bayes (NB). Genetic algorithms are used to find the right characteristics, while rough approaches for minimizing the size of the data for prediction are used. Decision rules are used as a success appraisal classification for the expected effects and grades. The hybrid approach to efficient extraction of functionality is provided in Kermani et al. [8] with the Genetic Algorithm with Neural Network (GANN). When the chosen functionality is efficient and important, the neural network performs better.

Jain [9] notes that data mining can be used to define, approximate, forecast, equate laws, cluster, and visualize activities. These activities may be predicted, categorized, and estimated by supervised learning categories which prepare a model based on the available data representing one or more attributes. Clustering is a major operation in these techniques which allows for data to be clustered based on the existence or a disease symptom. It may then be used for data cutting at the primary point. One of the basic and essential clustering algorithms is the K-means algorithm. Classification can be depending on parametrically, semi metrically, or nonparametrically. A sample from the known distribution is used by a parametric method, while a parameter used a sample from an uncertain distribution [10]. K-means clustering [11] is semi-parametric, and it is simpler to identify data sets assuming k clusters. K-means have the key benefit that if the number of clusters is tiny, it can be fast machine speed for the major element.

Bradley and Fayyad [11] used k-means for initial points refining and obtained a strong low running period. Mary and Raja [12] have used k-means algorithms to increase the cluster efficiency for refining groups and extended the optimization of ant colonies (ACO). In 2014, Wang et al., who can cluster both numerical and categorical results, was developing the clustering system of molecular regularized consensus patient stratification (MRCPS). Centered on the optimization process [13]. Besides, Rahideh and Shaheed [14] proposed a more precise, sensitive, and unique classification method focused on k-means and fluid c-mean. Methods for the classification of breast cancer K-means and c-means blurry were utilized, with improved recall, procedure time, and physiotherapy [15] achieved. Furthermore, Festa [16] also suggested a skewed random-key genetic data cluster, which is comparatively helpful than other similar approaches. Chen [17] suggested a highly effective, hybrid intelligent model for the collection of functions used for clinical cancer of the breast. Wei et al. [18] have suggested a new DNA sequence classification clustering algorithm and its interaction by utilizing a new clustering algorithm. Ahmad and Yusoff [19] have successfully developed a modern clustering algorithm k-means that can use

mixed numerical and categorical functions. Compared to other clustering algorithms, it was found to be successful.

There has been tremendous attention given to data clustering many apps, including data processing, extraction of records, image segmentation and the grading of the pattern. Extended knowledge volumes Clustering is quite clear from the advancements in technology Big data size is a daunting challenge [20]. To answer this issue, many researchers are attempting to develop successful algorithms for concurrent clustering. Zhao et al. [21] Suggest an algorithm focused on parallel k-means on the basic yet potent parallel programming Map Reduce technical. The authors focused on parallel implementation of k-means on particular data set. However, not tested with different performance metrics.

In the automatic diagnostic framework for breast cancer focused on the AR Association Rules approach, the data collectors used are Wisconsin Breast Cancer, the three-fold cross-validation procedure was used during a training and validation period and the findings from the experiment were carried out [22]. The same data collection used by Chaurasia et al. [23] where the authors suggested the prediction of benign and malignant ash cancer by using the Naive Bayes NB, the RBF Network and the J48 algorithms reveals that NB is the strongest predictor with 97.3 precision, while the RBF Network achieved 96.77% and the j48 algorithm achieves a 93.41% accuracy. The implementation of ML machine learning algorithms using the Wisconsin breast cancer data collection was submitted for the purposes of breast cancer diagnosis, for which six ML algorithms corresponding to GRU-S VM, Linear Regression, Multilayer Perceptron MLP, NN, Softmax Regression and Help Vector Machine SVM were submitted to the experiment [24].

A novel method is proposed to identify breast cancer using data mining techniques, with the intent of comparing three classification techniques utilising the Weka framework used in the usage of the algorithms SMO, IBK and BF Tree, the data set used corresponds to Breast Cancer Wisconsin [25]. A comparative analysis [26] was performed between the K-means and the FCM fuzzy C-means for breast cancer identification, which focuses on first evaluating the outputs of K-clustering algorithms and FCM, secondly the integration of different machine measures to boost the grouping accuracy of the above listed techniques, where FCM obtained b b A prediction analysis of breast-cancer recurrence using data mining technology has been presented [27, 28]; the study suggested the implementation of various classification algorithms such as C5.0, KNN, Naive Bays, SVM, and the experimental findings indicate that C5.0 has best results at 81.03% accuracy levels with K-Means, EM, PAM, Fuzzy C-means clustering process.

3. METHODOLOGY

The parallel k-means methodology aims to analyse the microarray breast cancer data-using a map-reduce-based k-means clustering algorithm. The methodology here is described as initially we represent the k-means algorithm and after represents methodology in two phases. One is a map with k-means and another one is a reducer. The details of the methodology with microarray breast cancer data are represented below. Here we used the idea of Zhao et al. [21] of the parallel k-means methodology and modified as a two-step model with the only map and reduce functions represented in Figure 2. Also, the modified k-means model is applied to the genomic breast cancer data set, and observed the performance of the model. Considered with different performance metrics of Silhouette coefficient, cluster formation time, and execution time.

3.1 K-means clustering algorithm

The K-means algorithm is the most common algorithm. It aims to find K clusters that reduce the amount of the Euclidean square distance between each sighting of the cluster and its mean. For its simplest form, the algorithm K-means alternates iteratively from one stage to the next: (1) allocate the cluster with the closest center in a given collection of cluster centers; and (2) change each cluster center as the mean sample of all points inside that cluster for a given assignment of observations. For Stage 1 the initial center values are often a random K sample. It converges usually through one of the various local and not the worldwide optimum. The more complex algorithms are likely to find the right locale. Regardless of the algorithm used, the algorithm should be begun continuously using various initial values, thereby improving the probability of reaching a successful optimum locally.

One of the most often employed methods for clustering is the k-means algorithm. It begins with the initialization of k cluster centers, in which k is specified. So the cluster with the score closest to each entity (input vector) of the dataset is allocated. The average mean of each cluster is determined such that the cluster core can be modified. This modification is attributable to improvements in each cluster's membership. Until no more adjustments are made, the procedures used for the assignment and modification of the cluster centers are replicated.

The following measures are used in the algorithm:

Collection of n data points: / Required: d=}, {d1, d2, d3, ..., di, ..., dn}

k / Amount of clusters you like

A collection of k clusters is given.

Steps: Stages:

1. Select k data points from D as initial centroids arbitrarily;

2. Repeat.

Assign per di point to the cluster that has the centroid closest;

For each cluster, measure the new mean;

Before the conditions for convergence are fulfilled.

The k-means algorithm does not always have an optimum setup, which correlates to the minimum global objective function, although proven to have been completed always. The algorithm is also adaptive to the original cluster centers randomly chosen. This impact can be minimized by running the k-means algorithm repeatedly. K-means is an easy and modified algorithm for several problem areas.

Here Figure 2 represents the working of the parallel process methodology for microarray breast cancer. The map-reduce k-means algorithm requires one sort of Map Reduce work, the microarray data is given. The map function executes a process in which any sample is allocated to the nearest center, while the reduction function updates the new centers. A combiner feature is built to resolve a partial combination of the intermediate values with the same key within a single map operation, to minimize the network connectivity costs. Map feature: The data collection is saved as a < key, value > pair sequence file in the HDFS [11], each of which contains a dataset record. The main is the byte offset of this document to

the data file starting point and the meaning is the contents string of that record. The dataset is separated and sent to all mappers worldwide. The distance measurements are then done in parallel. The parallel K-Means create a global variant center for each map task which contains details about cluster centers. Provided the details, the mapper will measure each sample's nearest center point. The intermediate values then consist of two parts: the center point index and the sample detail.

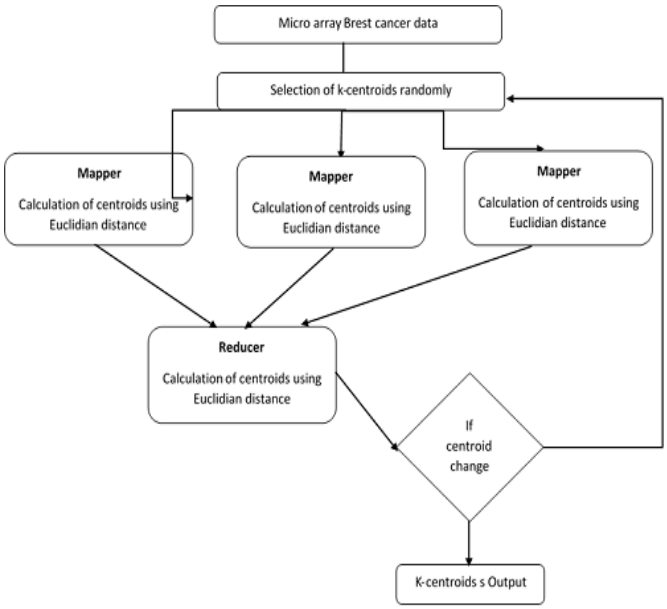

Figure 2. Methodology of parallel process for microarray breast cancer

Algorithm 1 displays the pseudocode of the map function.

Input: list of data points D = {d1, d2... dn}, set of initial selected centroids C = {c1, c2, , ck}

Output: output list (ol) consisting (Ci, Dj) pairs

Terminology: best Centroid (bC), current centroid (cc)

where $1 \leq i \leq n$ and $1 \leq j \leq k$

Process

1. M1 ← {d1, d2... dm}
2. cc ← C
3. Distance between (p, q) =√Σdi=(pi – qi)1/2 (Where pi (or qi) is p (or q) coordinate in scale I)
4. For every xi ε M1 such that $1 \leq i \leq m$ do
5. bC ←null
6. MinDist ← ∞
7. For every c ε cc do
8. dist← distance (xi, c)
9. if (bC = null || dist<minDist)
10. then
11. minDist←dist
12. bC ← c
13. end if
14. end for
15. Produce (bC, xi)
16. i=i+1
17. end for
18. return ol

We use a combiner to merge after each map task the same map task's intermediate information. Because the middle data are processed Local host disc, connectivity expenses cannot be consumed by the procedure. We apply the importance of the given points partially in the combined feature same cluster. To measure the average value of each item the number of samples in the same cluster should be registered in that cluster task for map. The data from which the reduction feature is entered is any host's combine feature. The mixture feature defines the data contains a partial sample total in the same cluster and sample number. We should add all the samples and measure the sum to minimise the function. Number of cluster samples allocated. We will therefore have the latest thing Next version of the centres.

Reducer algorithm

Input: pair of Key and Value;

Where key = best Centroid (bC) and Value = Objects assigned to the lpgr x centroid by the mapper

Output: pairs of Key and Value;

Where key = old Centroid (oC) and value = new Best Centroid (nbC) That is the new centroid (nc) value determined for the best centroid.

Terminology:

Output list (opl), new Centroid List (nCL), sum of objects (so), number of objects (no), centroid (ce)

1. Process
2. opl ← opl from mappers
3. u ← { }
4. nCL ← null
5. for all z opl do
6. ce ←z.key
7. obj ←z.value
8. ce ← obj
9. end for
10. for all ce ε u do
11. nc, so, so ← null
12. for all obj ε u [ce] do
13. so = so +obj
14. no = no +1
15. end for
16. nc ← (so + no)
17. produce (ce, nc)
18. end for
19. end

4. EXPERIMENTAL ANALYSIS

4.1 Data set

The data set comprising 54676 genes (columns) and 151 samples (rows) of gene expression values. In that dataset (column type"), there are five distinct types of breast cancer (plus healthy tissue). More details on this dataset and other file formats such as TAB and ARFF, data visualization, classification, and clustering references are accessible on the official CuMiDa web pages under the Id GSE45827: CuMiDa stands for providing more accurate data points for machine testing, separate from current databases for manual and carefully-cured sample accuracy, unnecessary samples, backgrounder correction, and normalization.

4.2 Results & discussion

Here the below results show that the map-reduce based k-means algorithm applied on breast cancer micro array data is compared with the existing k-means, K-NN, and SVM models.

The matrix of uncertainty is written to the right. In the matrix, every cell is a count of how many cases in a true class each of the projected classes has been divided. We will see with an uncertainty matrix whether a true class is confused. Figure 3 represents the Brest cancer research uncertainty matrix.

Silhouette relates to the way accuracy is represented and avoided within data clusters. This methodology offers a concise description of the classification of each item. The value of the silhouette indicates how an entity is identical to its cluster in contrast with other clusters. Figure 4 and Figure 5 reflect the coefficient silhouettes for the clustering of data on breast cancer concerning the multiple data set clusters.

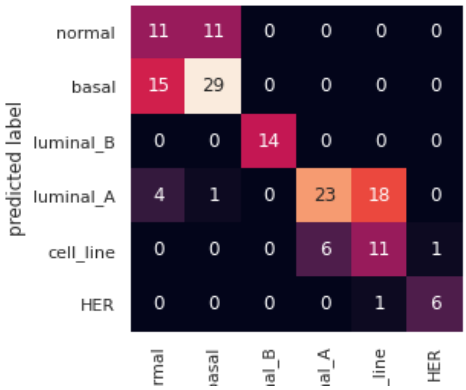

Figure 3. Confusion matrix for Brest cancer data analysis

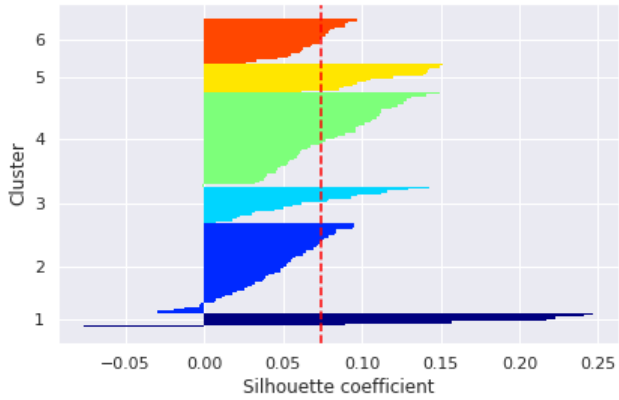

Figure 4. Silhouette coefficient for breast cancer data clustering

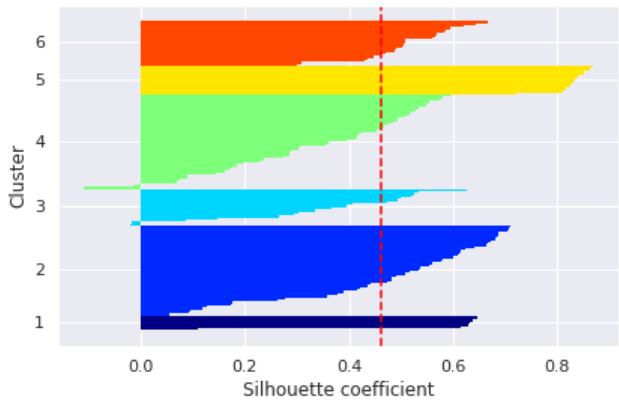

Figure 5. Silhouette coefficient for breast cancer data clustering

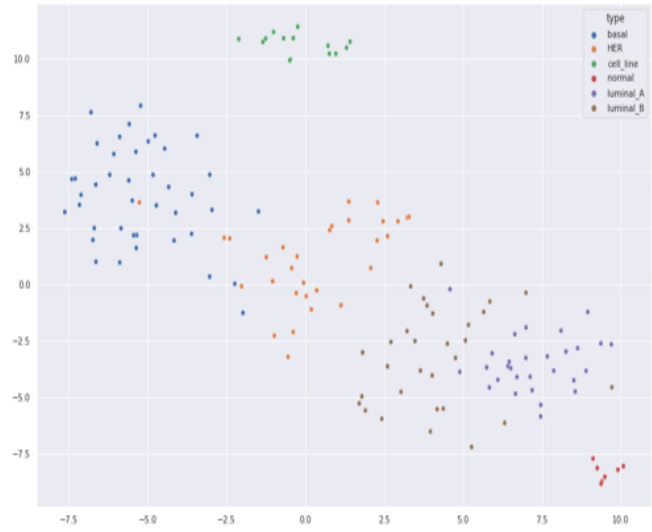

Figure 6. Clusters using parallel k-means methodology

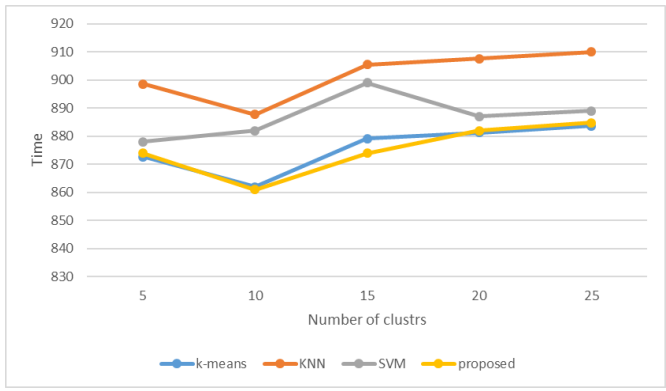

Figure 7. Cluster formation time

Here Figure 6 represents the clusters of different clusters formed by the parallel k-means methodology by taking the microarray data. Each relevant feature is formed as a cluster. The figure represents the different clusters with different color points.

Figure 7 represents the time for forming the clusters of different existing and parallel k-means methodology by giving the microarray breast cancer data. The cluster formation time is composed using a different number of cluster centers. The parallel k-means is fast concerning the number of clusters than all other existing models because of the parallel working of the methodology.

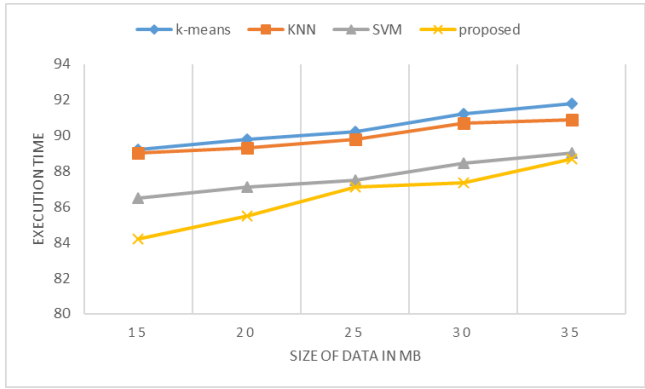

Figure 8. Execution time

Figure 8 represents the execution time for forming the clusters of different existing models and the parallel k-means methodology by giving the microarray breast cancer data. The cluster execution time means time taken for initialization, processing, and making clusters from the data set. The graph is drawn by varying the size of the input data and observed the time taken for executing the different existing models and parallel k-means models. The methodology is fast concerning the number of clusters than all other existing models because of the parallel working of the methodology.

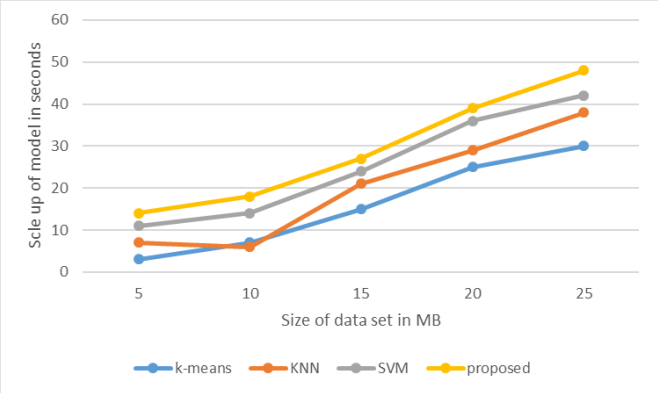

Figure 9. Scale-up in performance

The data scale-up assessment process (Figure 9) is used to test the efficiency of the parallel k-means. In our experiments on the data scale, the parallel k-means, KNN and SVM and, normal k-means for a fixed-size parallel process methodology cluster with different data sizes and the execution time for each experiment are reported. The subsequent experimentation period offers a basis for evaluating and examining output variations between current and parallel k-means methodology. Compared to current ones, a specifically parallel k-means methodology is best performed.

5. CONCLUSION

K-means clustering algorithm, used in this study for classification of microarray breast cancer dataset, is an unsupervised learning algorithm. The algorithm is unattended. The way to identify data sets with k clusters is quick and fast. In this analysis, an integrative cluster formulation of multi-variant parameters that has correctly evaluated the dataset for breast cancer was examined. The findings indicate, for the correct classification of the dataset, that Euclidean / Manhattan distances with the greatest difference and the same center of gravity are a safer alternative. This ensures that the model of our methodology will reliably and efficiently evaluate the micro collection of breast cancer datasets in k-means parallel processing maps. The comparative findings indicate that the parallel k-means methodology is successful.

REFERENCES

[1] Ahmad, L.G., Eshlaghy, A.T., Poorebrahimi, A., Ebrahimi, M., Razavi, A.R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health & Medical Informatics, 4(2): 1-3. https://doi.org/10.4172/2157-7420.1000124

[2] Aljawad, D.A., Alqahtani, E., Ghaidaa, A.K., Qamhan, N., Alghamdi, N., Alrashed, S., Alhiyafi, J., Olatunji, S.O. (2017). Breast cancer surgery survivability prediction using Bayesian network and support vector machines. 2017 International Conference on Informatics, Health & Technology (ICIHT), Riyadh, pp. 1-6. https://doi.org/10.1109/ICIHT.2017.7899000

[3] Asri, H., Mousannif, H., Al Moatassime, H., Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83: 1064-1069. https://doi.org/10.1016/j.procs.2016.04.224

[4] Bethapudi, P., SreenivasaReddy, E., Sitamahalakshmi, T. (2015). A computational approach for better classification of breast cancer using genetic algorithm. International Journal of Engineering and Computer Science, 4(6): 12853-12858.

[5] Devi, R.D.H., Devi, M.I. (2016). Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast.

[6] Elouedi, H., Meliani, W., Elouedi, Z., Amor, N.B. (2014). A hybrid approach based on decision trees and clustering for breast cancer classification. 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, pp. 226-231. https://doi.org/10.1109/SOCPAR.2014.7008010

[7] Elshazly, H.I., Ghali, N.I., Korany, A.M.E., Hassanien, A.E. (2012). Rough sets and genetic algorithms: A hybrid approach to breast cancer classification. 2012 World Congress on Information and Communication Technologies, Trivandrum, pp. 260-265. https://doi.org/10.1109/WICT.2012.6409085

[8] Kermani, B.G., White, M.W., Nagle, H.T. (1995). Feature extraction by genetic algorithms for neural networks in breast cancer classification. Proceedings of 17th International Conference of the Engineering in Medicine and Biology Society, Montreal, Quebec, Canada, pp. 831-832. https://doi.org/10.1109/IEMBS.1995.575385

[9] Jain, R. (2015). Introduction to data mining techniques. http://www.iasri.res.in/ebook/expertsystem/datamining.pdf., accessed on 22 April 2015.

[10] Alpaydin, E. (2014). Introduction to Machine Learning. MIT press, Cambridge, Massachusetts, United States.

[11] Bradley, P.S., Fayyad, U.M. (1998). Refining initial points for k-means clustering. In: Proceedings of the 15th International Conference on Machine Learning (ICML), Morgan Kaufmann, San Francisco, pp. 91-99.

[12] Mary, C., Raja, S.K. (2009). Refinement of clusters from k-means with ant colony optimization. Journal of Theoretical & Applied Information, 6(4): 28-32.

[13] Wang, C., Machiraju, R., Huang, K. (2014). Breast cancer patient stratification using a molecular regularized consensus clustering method. Methods, 67(3): 304-312. https://doi.org/10.1016/j.ymeth.2014.03.005

[14] Rahideh, A., Shaheed, M.H. (2011). Cancer classification using clustering based gene selection and artificial neural networks. The 2nd International Conference on Control, Instrumentation and Automation, Shiraz, pp. 1175-1180. https://doi.org/10.1109/ICCIAutom.2011.6356828

[15] Vanisri, D., Loganathan, C. (2010). Fuzzy pattern cluster scheme for breast cancer datasets. 2010 International Conference on Communication and Computational

Intelligence (INCOCCI), Erode, pp. 410-414
[16] Festa, P. (2013). A biased random-key genetic algorithm for data clustering. Mathematical Biosciences, 245(1): 76-85. https://doi.org/10.1016/j.mbs.2013.07.011
[17] Chen, C.H. (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Applied Soft Computing, 20: 4-14. https://doi.org/10.1016/j.asoc.2013.10.024
[18] Wei, D., Jiang, Q., Wei, Y., Wang, S. (2012). A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics, 13(1): 174. https://doi.org/10.1186/1471-2105-13-174
[19] Ahmad, F.K., Yusoff, N. (2013). Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier. 2013 13th International Conference on Intellient Systems Design and Applications, Bangi, pp. 121-125. https://doi.org/10.1109/ISDA.2013.6920720
[20] Bache, K., Lichman, M. (2013). UCI machine learning repository. 1990: 92. http://archive.ics.uci.edu/ml
[21] Zhao, W., Ma, H., He, Q. (2009). Parallel k-means clustering based on map reduce. In: Jaatun M.G., Zhao G., Rong C. (eds) Cloud Computing. CloudCom 2009. Lecture Notes in Computer Science, vol 5931. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-10665-1_71
[22] Karabatak, M., Ince, M.C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36(2): 3465-3469. https://doi.org/10.1016/j.eswa.2008.02.064
[23] Chaurasia, V., Pal, S., Tiwari, B.B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 12(2): 119-126. https://doi.org/10.1177/1748301818756225
[24] Agarap, A.F.M. (2018). On breast cancer detection: an application of machine learning algorithms on the Wiscons in diagnostic dataset. In: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, pp. 5-9. https://doi.org/10.1145/3184066.3184080
[25] Chaurasia, V., Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization), 2(1): 1-17.
[26] Dubey, A.K., Gupta, U., Jain, S. (2018). Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science Engineering and Information Technology, 8(1): 18-29. https://doi.org/10.18517/ijaseit.8.1.3490
[27] Ojha, U., Goel, S. (2017). A study on prediction of breast cancer recurrence using data mining techniques. 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, pp. 527-530. https://doi.org/10.1109/CONFLUENCE.2017.7943207
[28] Lichman, M. (2019). UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml/datasets/breast+cancer