Use Of Fpga For Real-Time K-Means Clustering Algorithm

Muhammed YILDIRIM*‡, Ahmet ÇINAR*

*‡ Fırat University, Department of Computer Engineering,muhyldrm23@gmail.com

**Fırat University, Department of Computer Engineering, acinar1972@gmail.com

‡ Corresponding Author; Muhammed YILDIRIM, +90 424 237 0000-6315 muhyldrm23@gmail.com

Received: 22.08.2019 Accepted: 26.09.2019

Abstract- Data mining is important methods in the data processing step. Due to the fact that computer technologies are becoming increasingly cheap and their power is increasing day by day, they allow computers to store data in larger quantities [1]. Owing to the improving of technology, many transactions are recorded in an electronic device and this records can be safely stored. This data can easily be accessed when requested. By means of developing technologies, it is ensured that these processes are getting more day by day at a lower cost. Therefore, it is of great importance to be able to process these data of high size. Clustering algorithms that aggregate the data in the database under groups or clusters to bring together objects with similar properties have a great deal of data mining proposition. In this paper, it is aimed to collect 2 clusters based on the similarities of 60 data obtained from 2 different wheat varieties using k-means clustering algorithm based on Fpga architecture. Since the FPGA architecture has the ability to perform parallel processing, it will shorten the processing time and so efficiency will increase. Also, the ability to use FPGA's over and over again provides an extra advantage. The proposed system is designed using the verilog hardware identification language on the DE2_115 Fpga board.

Keywords: Data Mining, Fpga, K-means, Verilog

1. Introduction

Together with the developing technology, the amount of data held in the databases increases day by day. New Technologies need to be developed in order to process these data more easily.Data mining is the process of processing high amounts of data held in databases, obtainig meaningful data and making it ready for use. In short, it can be said that it covers all data analysis techniques for data mining process[2]. Data mining is a method that uses the available data to acquire the knowledge necessary to solve the present problem or make predictions about the future. It is also a way of revealing patterns and rules hidden in databases [3].

Different techniques can be used in data mining. These techniques can be examined under two main headings: Descriptive and Predictive techniques [1]. In predictive techniques, a technique is developed using the known data, and this technique is used to forecast the result values for data sets whose results are unknown [1]. Factors and conclusions for predicting possible outcomes are included in the modeling. How much well estimated results are estimated is as important as the estimated result. In descriptive models, the goal of functions is not to predict a specific goal. The aim is to find relationships, connections and behaviors on the data. It aims to make determinations about behavior patterns using existing FPGA are manufactured by more than one company. But in general they work within the same logic. FPGAs are data then to define properties of sub data sets that show this behavior [4]. It is possible to examine the methods used in the data mining in three different titles.

- Classification and Regression
- Clustering
- Association Rules [3].

Since data mining runs on large data heaps, the transaction time is of great importance. Field Programmable Gate Array (FPGA) integration is used in this study to reduce the processing time. The biggest feature that separates the FPGA from other integrations is that it can be changed the internal configuration as we wish.

Another important feature of FPGA is the ability to perform parallel operations. Parallel processing means that you can make more than more operation at the same time. Ordinary integrators can either implement a very limited number of operations or not at all. FPGA can execute more than one operation at the same time according to the application type. This means that you can complete the process more quickly. This feature makes FPGA popular in applications where parallel processing is required.

programmed in so-called HDL languages. In this study, 2 different kinds of wheat clusters are separated using FPGA Architecture.

2. Method

There are various models and algorithms developed for data mining [5]. In this study, k-means clustering algorithm is used for grouping similar data. The application is implemented on verilog language using k-means clustering algorithm.

2.1. Clustering

Clustering is called as clustering or segmentation. The clustering method is interested in the descriptive model category in the data mining. Because the task of functions here is not a predictor of a specific result. In this method, relations between the data sets are revealed [6]. The goal is to obtain homogeneous subgroups of data from a heterogeneous set of data by separating the clusters of similar elements. That is to say, in the clustering method, similar data are gathered up under the same clusters. In the clustering method, it is aimed that the similarity rate among the clusters is at maximum level and the similarity rate between the clusters is minimum. In some cases, the clustering model can be used in advance of the classification model. The main difference separating the clustering function from the clustering function is that the cluster does not use the predefined data. There are no pre-defined data and examples in clustering functions [7], while the data defined for the classification function and their pre-existing values constitute the basic model. Figure 1 shows two data groups separated into a set A and a set B of data.

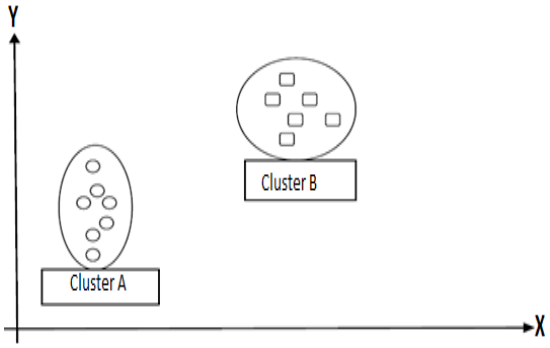

Fig. 1. Data Group A and B

Many of the clustering methods use metrics distances between data. Among these, the most commonly used are Euclidean metrics distance, Manhattan metrics distance and Minkowski metrics distance.

The models used in the clustering process can be classified as shown in Figure 2.

Clustering Models

1. Hierarchical Models
2. Partitioning Models
3. Density Based Models
4. Model Based Models
5. Grid-Based Models

Fig. 2. Clustering Models

2.1.1. K-Means Algorithm

The K-Means algorithm developed by J. Bacqueen in 1967 is one of the oldest known clustering algorithms. According to the K-means algorithm, each data can only be included in a cluster. It is an algorithm built on the idea of representing the central point of the cluster. The K-means algorithm divides n objects into k sub-sets with a maximum similarity rate within the cluster, According to a minimum similarity rate between the clusters [8]. The algorithm consists of the following steps.

1. K objects are chosen to indicate the centre point of each cluster or their averages. Since the selection of the initial cluster centers affects the result of the k-means algorithm, k objects can be chosen in the following form.
   - k pieces of random data are selected and assigned as the centers of the clusters.
   - The data are distributed randomly to k clusters and the centers of the clusters at the beginning are determined by taking the averages of the clusters.
   - The data with the extreme values are selected as cluster centers.
   - The closest points to the center of the dataset are taken as starting points.
2. The other objects outside the k objects are added to the closest similar clusters, taking into calculating the distances of the clusters to their average values. Euclidean distance is often used for distance calculations in general.
3. The averages of the cluster elements are taken for each cluster. This obtained average value is the new center point.
4. Again the distances to the center point of each data are calculated and added to the set of the closest center point. Averages of cluster elements are taken and new center points are found. In the clustering process, this process is repeated until the next step in the same way [9].

The easy implementation of the K-means algorithm gives a great advantage to work quickly on large quantities of data. Since the FPGA architecture has the ability to perform parallel processing, it will shorten the processing time and produce more efficient results.

3. Application and Result

In this chapter at first, the steps of the data mining process are discussed. The process of data mining is as follows.

• Understanding the business and business environment

• Verbal understanding

• Preparing the data

• Modelling

• Evaluation

• Spreading.

In this study, length and width data of 2 different wheat strains are taken from UCI Machine Learning Repository web site. Wheat can be shown as figure 3.

Fig. 3. a) Kama Wheat

Fig. 3. b) Rosa Wheat

The data preparation procedures, which are standard steps of the data mining, are applied. Noisy and incomplete data is completed so that the data can be synthesized easily on FPGA. After the data are prepared and saved in the text file, the application is performed in the verilog hardware description language using the quartus program. In practice, the data is first read from the text file and transferred to array. Then 2 initial values are selected and k-means algorithm steps are applied. As a result, the data are clustered in the same cluster with a maximum similarity ratio and a minimum cluster-to-cluster similarity ratio.

The application is implemented on verilog, which is one of the hardware identification languages. The software is implemented in quartus and simulated in modelsim program[10].

The following code block is used for reading files in Verilog program. Each always-block has the ability to be executed at the same time. In Always block, readmemb function is transferred from wheat.txt file to datar directory. Then the x and y coordinates of the data in the datar array are transferred to the datac array.

```
always @(posedge…)

initial $readmemb("wheat.txt",datar);

for (k=0;k<120;k=k+1) begin

        datac[l][0] =datar[k];

        datac[l][1] =datar[k+1];

        ……

end
```

The data read from the text file is parsed into clusters in the following always block which processes at the same time.

```
always @(…..)

while (center1[0][0]!=n[0][0] ) begin

  for (i=0;i<60;i=i+1) begin

        x= ((datac[i][0] - center1[0][0]) * (datac[i][0] - center1[0][0]))+((datac[i][1] - center1[0][1])*(datac[i][1] - center1[0][1]));

        y= ((datac[i][0] - center2[0][0]) * (datac[i][0] - center2[0][0])) +((datac[i][1] - center2[0][1]) * (datac[i][1] - center2[0][1]));

        if(x<y) begin

        kama[a][0]=datac[i][0];

        kama[a][1]=datac[i][1];

        k1xtop=k1xtop + kama[a][0];

        k1ytop= k1ytop + kama[a][1];

        …….

        end

        else begin

        rosa[b][0]=datac[i][0];

        rosa[b][1]=datac[i][1];

        ……

        end

 end

        center1[0][0]= k1xtop / (a);
```

```
        center1[0][1]= k1ytop / (a);

        ……..
end
```

The Flow diagram of the application is shown in figure 4.

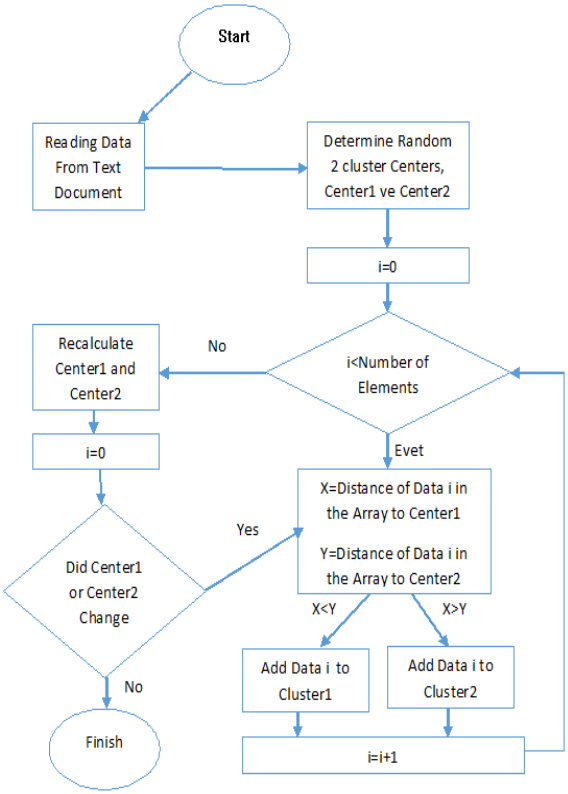

Fig. 4. Flow Diagram

Since the Verilog hardware description has the ability to perform parallel operations in the mirror, it is parsed into sets in another always block while read from the given file in an always block. This is a great contribution to the completion of the program in a shorter period.

After reading the files, the steps of the k-means algorithm are applied into Kama and Rosa data, Two different types of wheat are separated into two different clusters. When the K-means algorithm is applied, initial values for two clusters are initially selected. Then, as the program started to work, the center points of 2 coins are recalculated and the values of the center points are started to change. This process continues until the cluster's center points changed. After the center points have a fixed value, the program is terminated. After the program are finalized, Kama and Rosa wheat are separated from each other and collected in two different clusters.

The simulated image in the modelsim program of the obtained clusters and the data read from the text document is the same as in figure 5.

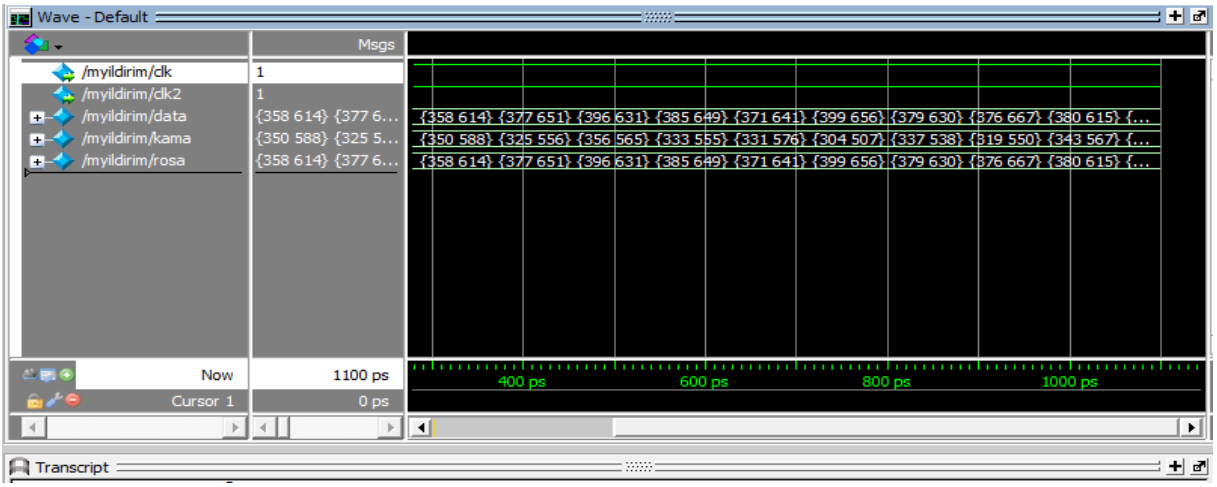

Fig. 5. Simulation View

The data obtained from the text file is transferred to the data sequence as seen in the simulation. The data array is defined in two dimensions to hold the width and length values. For example, (358, 614) has a width of 358 and 614 is a length.

The data in the data sequence are then divided into 2 clusters with a minimum similarity for kama and rosa by applying the k-means clustering algorithm.The width and length beginning data read from the text document are shown in Figure 6.

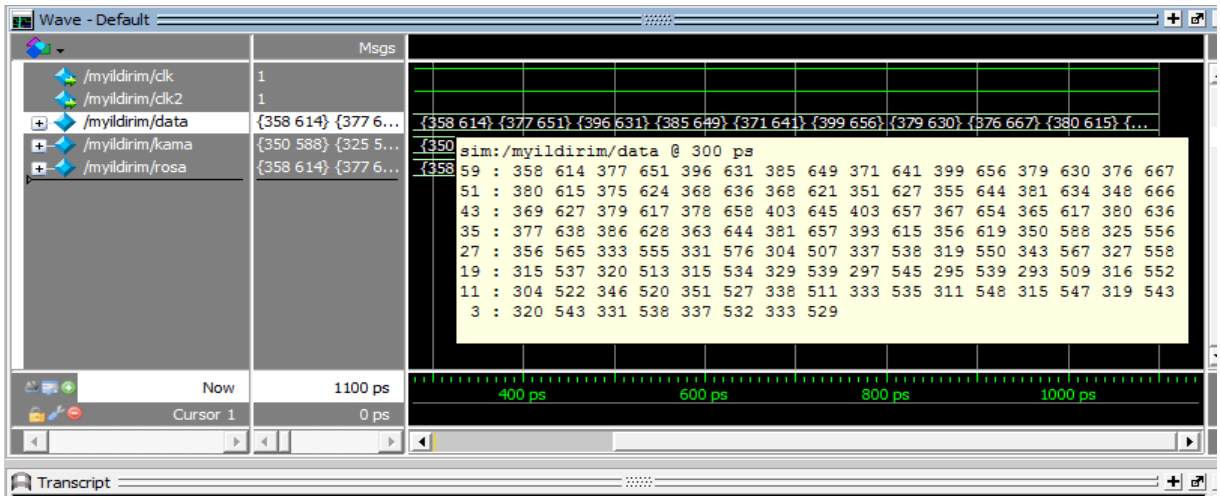

Fig. 6. Startup Data Read from Text File

The Kama values obtained after applying the K-means algorithm are shown in Figure 7.

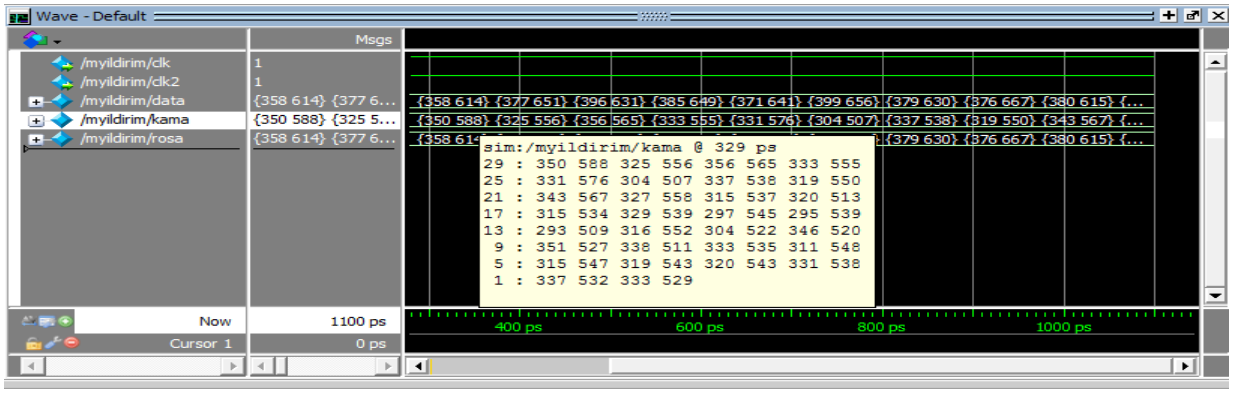

Fig 7. Kama Values

The Rosa values obtained after applying the K-means algorithm are shown in Figure 8.

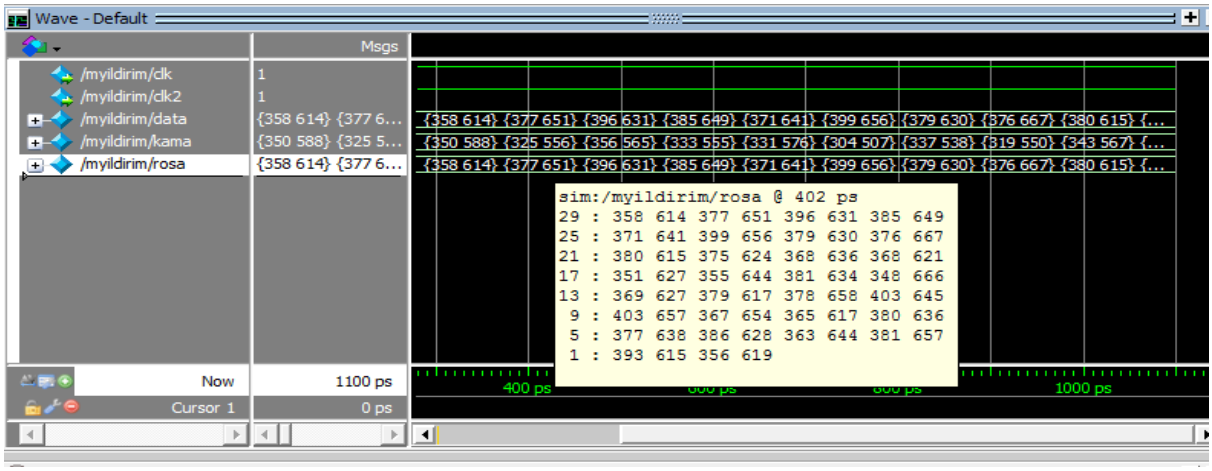

Fig 8. Rosa Values

Figure 6, Figure 7, Figure 8 shows the reading of the data at 402 ps. Since the Fpga architecture processes are parallel,

the times in Figure 6 and Figure 7 are realized within these times. In large data sets, this means that the transaction time is short.

In this study, Alteranın DE2_115 fpga board used. The DE2_115 board, connectors and board components are as shown figure 9.

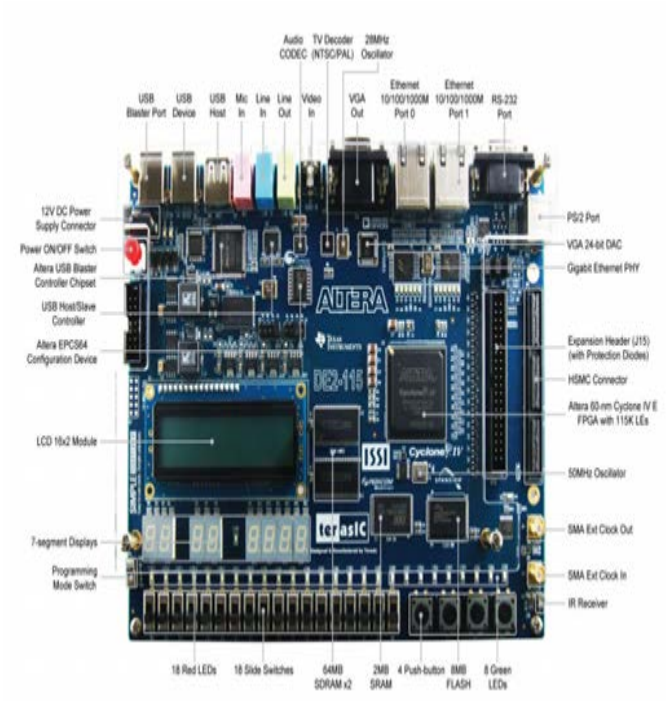

Fig. 9. The DE2_115 board.

We can adapt the DE2_115 FPGA board according to the k-means clustering algorithm. Due to the

reproducibility of the fpga board, the DE2_115 Fpga board is available for repeated use. In this way, the software can be developed on the same board.

The block diagram of the fpga board is as shown in Fig. 10.

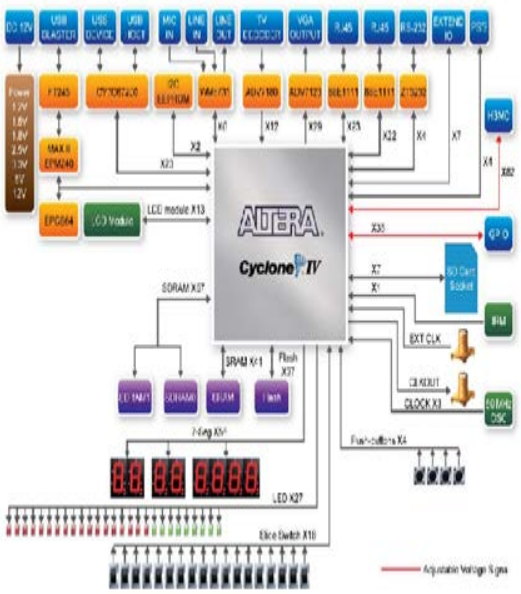

Fig. 10. Block Diagram of DE2_115 FPGA Board.

This study was carried out using quartus and modeling program. The flow chart of the software carried out in the Quartus program is as in figure 11.

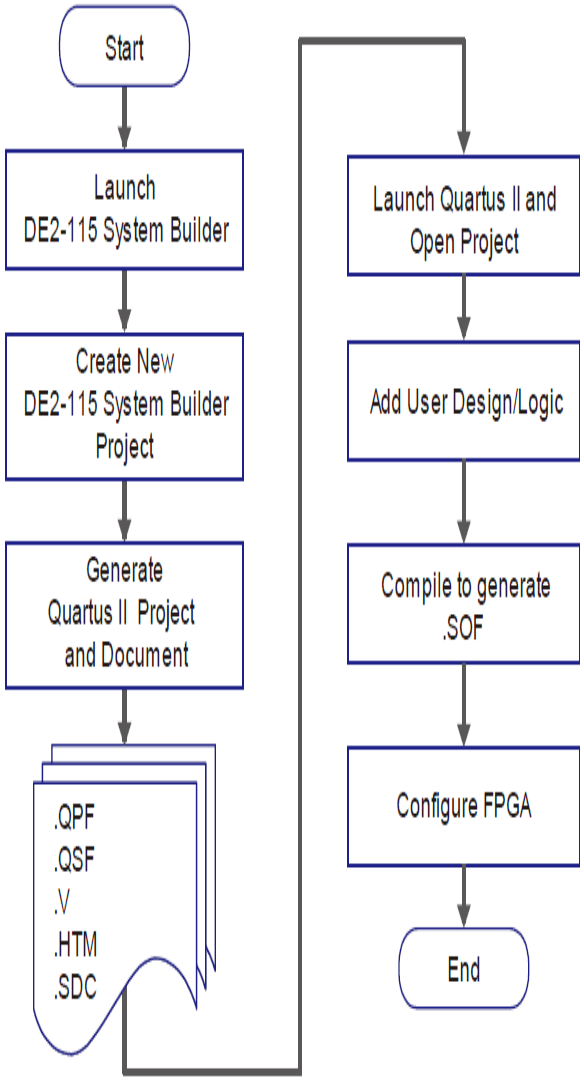

Fig. 11. The flow chart of the software carried out in the Quartus program

4. Conclusion

Normal integrators can not perform parallel operations at all or can perform a very limited number of operations. Fpga can imply a lot of operations parallel to each other at the same time.

In this application, the K-means algorithm is used in clustering algorithms which is one of the methods of data mining. Both k-means algorithm is an efficient algorithm and implementation on DE2_115 FPGA board produces efficient and fast results.

The K-means clustering algorithm is generally accepted and is used in many places .In the K-means clustering methods, the distances of each point in the priority data set relative to the center points of the clusters are calculated. Then the point is placed where the cluster belongs. Calculation of distances is a time-consuming work especially for big data sets and big numbers of clusters. In order to obtain a high performance, we need to decrease the calculation time of each point. Therefore, this calculation can be performed more quickly using FPGA architecture.

It is aimed to be the minimum intra-cluster similarity ratio in clustering process and to be the maximum between the clusters. Clustering analysis has been developed to elaborate the classification of individuals or objects. The different 2 types of wheat data in the hand are separated using the k-means algorithm. As a result, Kama and Rosa wheat varieties are collected in 2 different clusters.

References

[1] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

[2] Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49.

[3]Larose, D. T. (2014). Discovering knowledge in data: an introduction to data mining. John Wiley & Sons.

[4] Nadiammai, G. V., & Hemalatha, M. (2014). Effective approach toward Intrusion Detection System using data mining techniques. Egyptian Informatics Journal, 15(1), 37-50.

[5] Roiger, R. J. (2017). Data mining: a tutorial-based primer. CRC Press.

[6] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[7] Jayakameswaraiah, M., Babu, M. M. V., Ramakrishna, S., & Yamuna, M. P. (2016). Computation Accuracy of Hierarchical and Expectation Maximization Clustering Algorithms for the Improvement of Data Mining System.

[8] Stevens, R., Casillas, A., 2006. Artificial neural networks. Automated Scoring of Complex Tasks in Computer Based Testing: An Introduction. Lawrence Erlbaum, Mahwah, NJ, 259-312.

[9] Cui, X., Zhu, P., Yang, X., Li, K., & Ji, C. (2014). Optimized big data K-means clustering using MapReduce. The Journal of Supercomputing, 70(3), 1249-1259.

[10] http://dl.altera.com/?edition=lite