PERFORMANCE ANALYSIS OF DATA MINING TECHNIQUES FOR REAL TIME APPLICATIONS

Pradeep Nagendra Hegde[1], Ayush Arunachalam[2], Madhusudan H C[2], Ngabo Muzusangabo Joseph[1] , Shilpa Ankalaki[3], Dr. Jharna Majumdar[4, 5]
[1]BE Final Year Student, Dept. of Computer Science & Engg.
[2]BE Final Year Student, Dept. of Electronics & Communications Engg
[3]Asst Prof. Dept. of Mtech Computer Science & Engg.
[4]Dean R&D Prof. and Head Dept. of MTech Computer Science & Engg
[5]Head, Center for Robotics Research
Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore

Abstract
Data Mining plays a pivotal role in genomics, bio informatics, agriculture, fraud detection and financial banking. Extensive research has been done in this domain in the last three decades and numerous algorithms have been developed. This paper aims to make a comparative analysis of a few popular algorithms like Fuzzy C Means and K-Means for different data in real time and of development boards like Raspberry Pi 2, Udoo Quad. Conventionally, the value of K in K-Means is not known initially and is given as input. This paper aims to avoid manual input by the user by using Batchelor Wilkins' algorithm and Elbow Method. The performance of the data mining algorithms is determined using a set of Quality Metric (QM) parameters. The results show the most suitable algorithm for clustering for different databases. Timing analysis is also performed to determine the best suited development board for real time applications. The algorithms have been implemented in C on Udoo Quad and in Python on Raspberry Pi 2.
Index Terms: Batchelor Wilkins', Clustering, Data Mining, Elbow Method, Fuzzy C Means, K-Means, Raspberry Pi 2, QM Parameters, Udoo Quad

I. INTRODUCTION

Clustering is the task of splitting up data points into a number of groups such that data points belonging to a group are *closely correlated* than those in the other groups.

The objective is to *segregate* groups with *similar characteristics* and assign them into clusters. This project involves classifying the objects based on their features, derived from various methods. The features are subjected to *Data Mining* techniques like *Fuzzy C Means* and *K-Means* for performance comparison and classification. The efficiency of Data Mining techniques may vary with the input data set. Hence, choosing the most suitable Data Mining technique for a particular data set or purpose is vital to perform clustering.

This paper fulfills this prerequisite by presenting a *comparative analysis* on the aforementioned Data Mining techniques. The performances of these algorithms are compared using various *quality metric* parameters, namely *sum of squared error, intra cluster distance, and inter cluster distance.*

This paper also contributes to the *optimization* of K-Means clustering by *determining the value of K automatically using Batchelor Wilkins' algorithm*. This alleviates the burden of providing a random value for K.

II. LITERATURE SURVEY

Data mining deals with extracting information from a data set and transforming it into an understandable structure for further use. Clustering is the part of data mining which deals

with batching objects that are similar in some sense to one another. In the last five decades, extensive research has been carried out and many algorithms have been developed.

A. FUZZY C MEANS TECHNIQUE

Fuzzy C Means [2] clustering is a soft clustering algorithm. It allocates membership to each data point corresponding to each cluster center on the basis of distance between the data point and the cluster center. Lesser the separation between the data point and the particular cluster center, more is the membership of the data point with that particular cluster center.

The membership value is calculated between the data point and the cluster center. It is contrived as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(d_{ij}/d_{ik}\right)^{(2/m-1)}}$$

Where, μ_{ij} describes membership of i^{th} data to j^{th} cluster center and d_{ij} describes Euclidean distance between i^{th} data and j^{th} cluster center.

Select 'c' cluster centers and calculate the fuzzy membership 'μ_{ij}' using the above mentioned formula. After all the operations compute the fuzzy center 'v_j' using the earlier results.

$$v_j = \frac{\sum_{l=1}^{n}[(\mu_{lj})^m X\, x_l\,]}{\sum_{l=1}^{n}(\mu_{ij})^m}$$

Where, v_j describes j^{th} cluster center and x_i will be the data point values.

B. K-MEANS TECHNIQUE

K-Means [2] Clustering is a hard clustering algorithm, originally from signal processing, that is widely used for cluster analysis in Data Mining. K-Means clustering focuses on segregating n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as the template of the cluster.

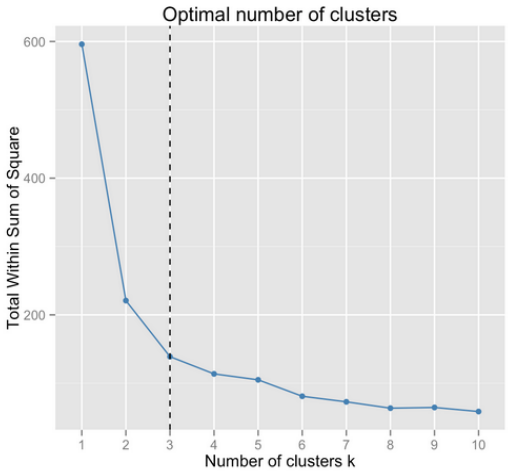
The problem is computationally difficult (NP-hard); however, there are high performance heuristic algorithms that are commonly utilized to ensure convergence to a local optimum occurs rapidly. These are usually similar to the expectation maximization (EM) algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Both employ cluster centers for data modelling. K-Means clustering tends to determine clusters based on their compactness or proximity among data points, while the EM mechanism allows clusters to have variegated structures and outlines.

$|x_i - v_j|$ describes Euclidean distance between i^{th} data and j^{th} cluster center. Based on this the clusters will be formed.

C. BATCHELOR WILKINS' TECHNIQUE

Batchelor Wilkins' [3] algorithm is a clustering algorithm used when number of cluster centers is unknown. It is a simple heuristic procedure for clustering and is sometimes referred to as **maximum distance algorithm**.

It differs from Fuzzy C means algorithm in the fact that the value of k i.e. the number of cluster centers isn't predefined. We determine the number of cluster centers automatically using Batchelor Wilkins' algorithm.



$$\textit{Centroid} = \frac{\sum_{i=0}^{m-1} f(x_i)}{m}$$

Where, $f(x_i)$ is the value of the data point x_i i.e. a feature or property and m is the number of such data points.

D. ELBOW METHOD

Elbow method is a technique used to determine the best value of k in K-Means and the best initial value of clusters in Fuzzy C Means algorithms. It is a technique used to interpret and validate within-cluster consistency.

E. QUALITY METRIC PARAMETERS

i. Sum of Square Error (SSE)

SSE [5] computes the average of the squares of the errors i.e. the difference between the approximator and the approximated values.

$$SSE = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i^2)$$

Where, x_i is the data point.

ii. Intra Cluster Distance (ICD)

Intra Cluster Distance [4] will compute the value between the data points and the supreme value of the data point within that cluster.

$$ICD = \sum_{i=1}^{n}(x_j - x_{max})^2$$

iii. Inter Cluster Distance (InCD)

Inter Cluster Distance [4] is the quality metric value which is calculated between the supreme data point of one cluster and the minimum data point of the adjacent cluster.

$$InCD = \sum_{i=1}^{n}(x_{max_i} - x_{min_{i+1}})^2$$

III. DEVELOPMENT BOARDS

This paper deals with implementation of Data Mining techniques on hardware for real time applications. Development boards like Raspberry Pi 2 and Udoo Quad are used in order to obtain results in real time.

Raspberry Pi 2

Raspberry Pi 2 is a low cost minicomputer which supports Linux and Windows 10 OS. It is used to interface sensors, actuators. It has 1 GB of RAM, 700 MHz ARM processor, 26 GPIO pins, Wi-FI module, and provisions for USB, Ethernet, HDMI, Audio, and Video. It is extensively used for IoT, security, and gaming applications.

Udoo Quad

Udoo Quad is a powerful single board computer which supports Linux and Android OS. It is a quad core processor which provides powerful prototyping. It has 1 GB of RAM, ARM Cortex iMX 6+ Arduino Due processors, 72 GPIO pins, Wi-FI module, and provisions for USB, HDMI, Ethernet, Audio, and Video. It is used for real time image and video processing.

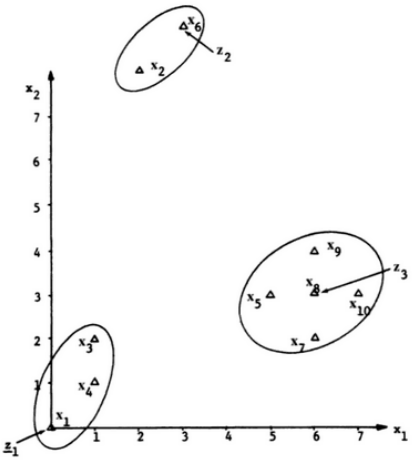
IV. PROPOSED WORK

Data Mining algorithms like Fuzzy C Means, K-Means are implemented to cluster the input data points.

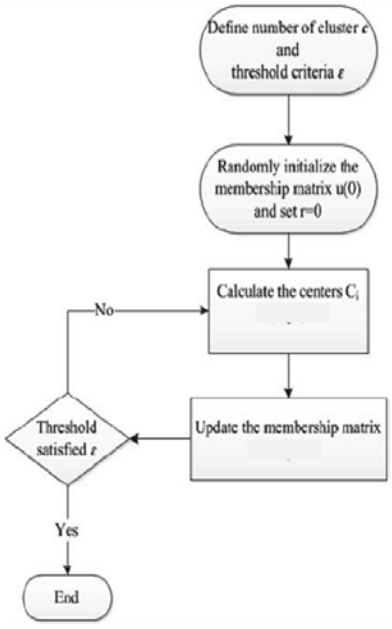
Batchelor Wilkins' algorithm is used to determine the initial number of clusters i.e. K automatically. Elbow Method is used to find the best value of k among those specified by the user.

Both Fuzzy C Means and K-Means algorithms are implemented on development boards like Raspberry Pi 2 and Udoo Quad.

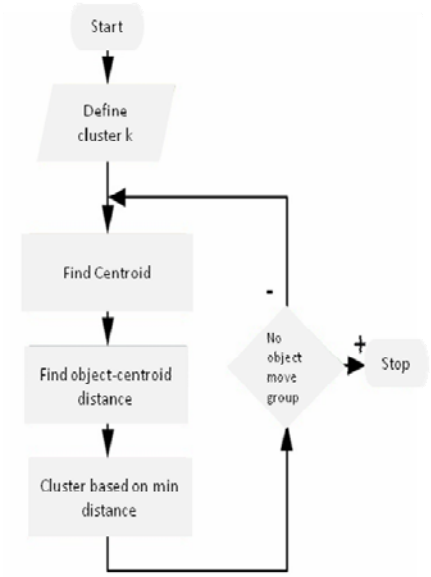
Batchelor Wilkins' sample



Fuzzy C Means flowchart



K-Means flowchart



V. EXPERIMENTAL RESULTS

The above discussed Data Mining algorithms are compared with their respective results and QM analysis for different data.

Before analyzing the results of the different Data Mining algorithms, it's necessary to study the trends of the quality metrics which represents the performance of the Data Mining algorithms. Fig.1 shows the trends of the quality metrics parameters.

Fig 1. Trends of Quality Metrics Parameters

Quality Metrics	Characteristics
SSE	Higher Value: Poor Clustering Lower Value: Good Clustering
ICD	Higher Value: Poor Clustering Lower Value: Good Clustering
InCD	Higher Value: Good Clustering Lower Value: Poor Clustering

Fig 2. QM Analysis of Agriculture Data

Algorithm	Quality Metrics Parameters		
	SSE	ICD	InCD
Fuzzy C Means	1,387,570.816	596,285.875	71,804,410
K-Means	**596,285.9**	**689,432**	**87,891,234**

Fig 3. QM Analysis of Leaf Images Data

Algorithm	Quality Metrics Parameters		
	SSE	ICD	InCD
Fuzzy C Means	50,397.87109	8,763.743164	42,285.863281
K-Means	**12,405.314453**	**6,405.314453**	**258,815.375**

Fig 4. Timing Analysis (in ms) of Agriculture Data

Algorithm	Device	
	Raspberry Pi 2	Udoo Quad
Fuzzy C Means	3.57	2.031
K-Means	**0.457**	**0.2320**

Fig 5. Timing Analysis (in ms) of Leaf Images Data

Algorithm	Device	
	Raspberry Pi 2	Udoo Quad
Fuzzy C Means	2.497	2.051
K-Means	**0.391**	**0.221**

Fig 6. Clustering of Leaf Images Data using K-Means

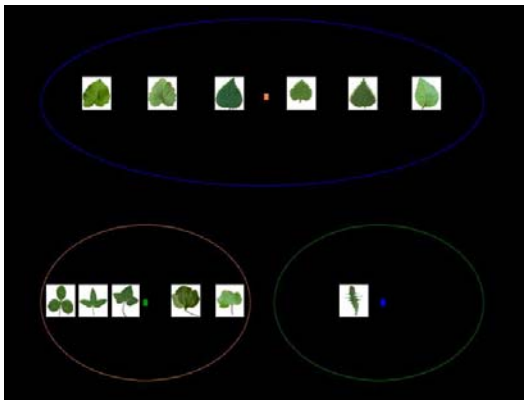
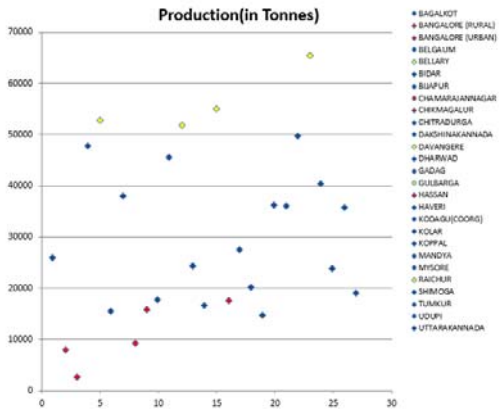


Fig 7. Clustering of Agriculture Data using K-Means



VI. CONCLUSION

This paper presents a comparative analysis of the performance of different data mining algorithms namely, Fuzzy C Means and K-Means. The performance has been compared using various QM parameters namely, sum of squared error, intra cluster distance and inter cluster distance. This paper also presents a comparative analysis of the performance of Raspberry Pi 2 and Udoo Quad development boards by performing timing analysis. From the above results, this paper concludes that K-Means algorithm provides the best clustering results for agriculture data and leaf images data and that Udoo Quad is the better development board.

The paper also presents a novel procedure for determining the initial number of clusters, K for performing K-Means clustering which was computed with the help of Batchelor Wilkins' algorithm.

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to Prof. N. R Shetty, Advisor and Dr. H C Nagaraj, Principal, Nitte Meenakshi Institute of Technology for giving constant encouragement and support to carry out research at NMIT.

The authors extend their thanks to Vision Group on Science and Technology (VGST), Government of Karnataka to acknowledge our research and providing financial support to setup the infrastructure required to carry out the research.

REFERENCES

[1] Gonzalez, R. C., & Woods, R. E. (2002). Digital image processing (2nd Ed.).NJ: Prentice Hall.

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Addison-Wesley, 2005

[3] Greg Hamerly, Charles Elkan {ghamerly, elkan} (2003). Learning the k in k-means.

[4] Kirkland, Oliver and De La Iglesia, Beatriz (2013) *Experimental evaluation of cluster quality measures.*

[5] Ujjwal Maulik, Member, IEEE, and Sanghamitra Bandyopadhyay, Member, IEEE (2002). Performance Evaluation of Some Clustering Algorithms and Validity Indices.