Design of Scalable Machine Learning Algorithm for Handling Genomics Data

Preeti Jha, Aruna Tiwari

Indian Institute of Technology, Indore

phd1801201006@iiti.ac.in, artiwari@iiti.ac.in

Abstract

A major challenge in analyzing the data from high throughput genomics is the way to deal with the huge amount of data using a variety of traditional tools. The role of technology in genomics focus on the massive growth in genome sequencing, which has a development rate quicker than expected by Moore's law. The size of the data set is getting so much larger which requires huge data processing technology. In order to support the investigations in bioinformatics, explicitly on genomic variations and population genetics, we have proposed a feature extraction approach for converting DNA sequences and SNP sequences into the feature vectors. The main aim is to process such data so that it can be made usable to be provided as input to the machine-learning algorithm. The feature extraction approach extracts six relevant features corresponding to each sequence. We have also implemented the scalable kernel fuzzy clustering algorithms on Apache Spark, which perform the efficient clustering of Big Data due to their in-memory cluster computing technique. Our scalable kernel fuzzy clustering algorithm provides the better potential for the genomics data handling.

Proposed Kernelized Scalable Random Sampling with Iterative Optimization Fuzzy c-Means (KSRSIO-FCM)

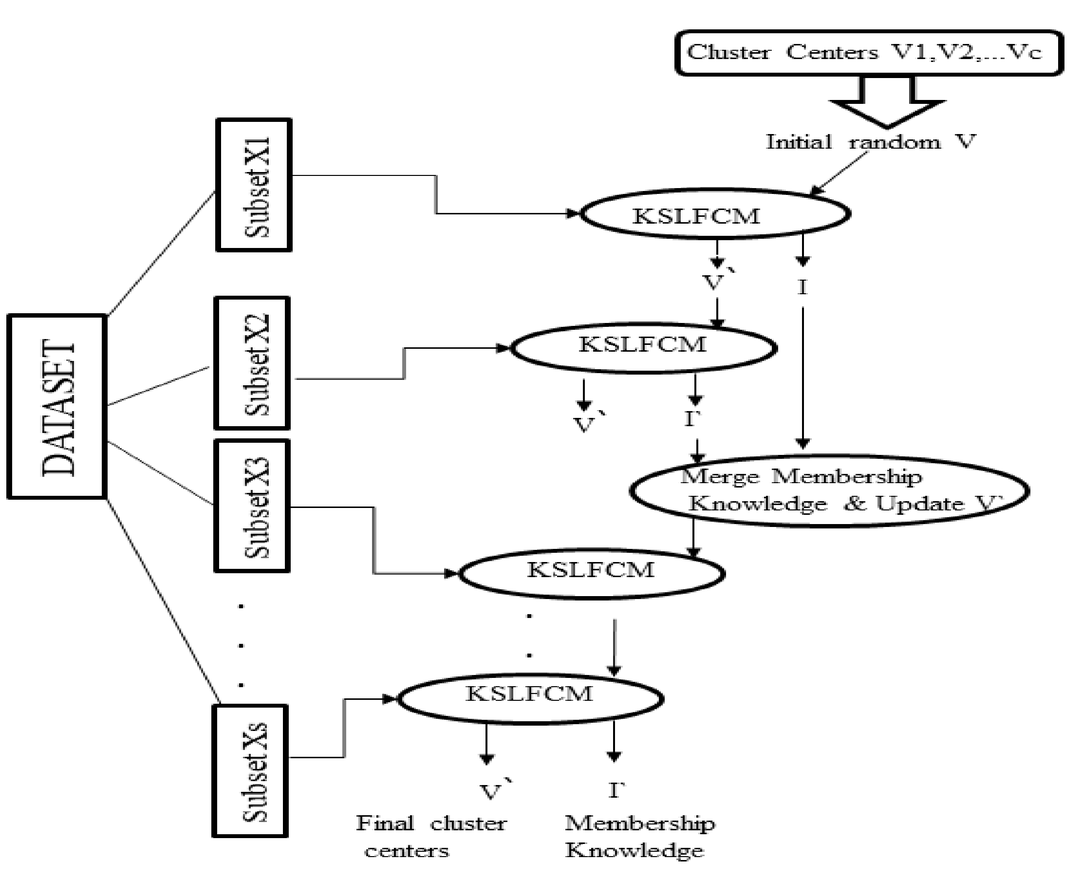

Fig. 1 Workflow of KSRSIO-FCM

Working of KSLFCM on APACHE SPARK

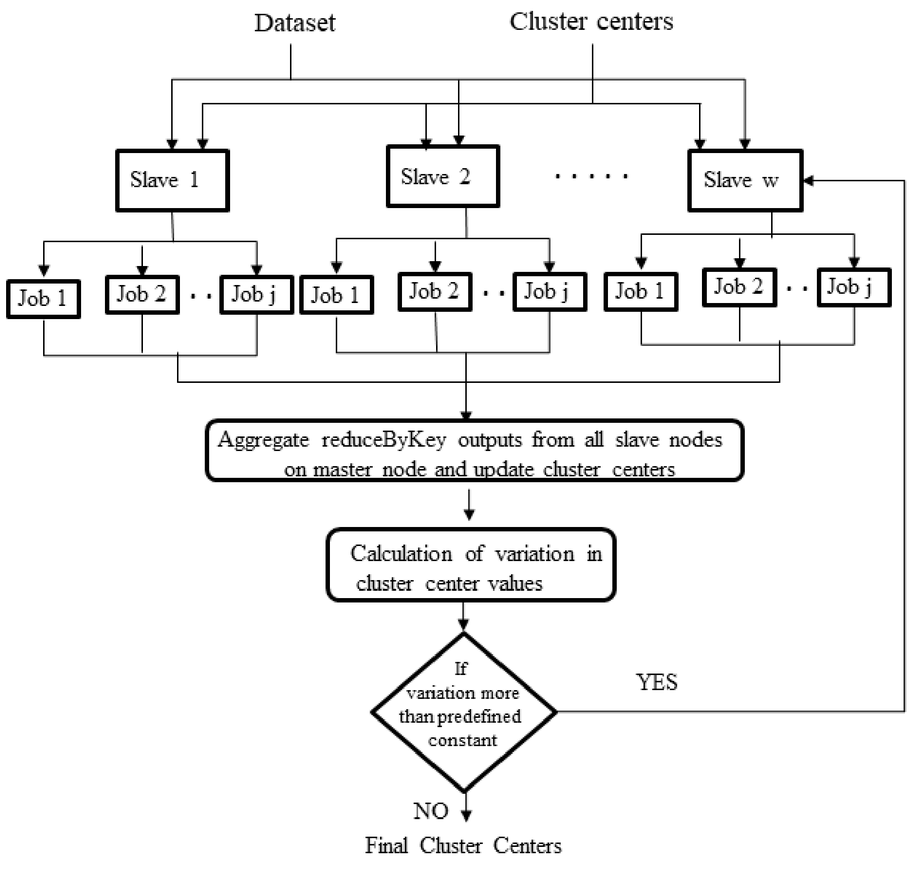

Fig. 2 Flowchart showing working of KSLFCM on APACHE SPARK

Steps of DNA and SNP Data Preprocessing

We started with the approach given in the paper [3]. In this paper, an algorithm is proposed to convert DNA data into 12 features, in which three features represent each nucleotide.

1)The nucleotide A, T, G, and C content from the DNA sequence are chosen as the first parameter in the vector analysis. Therefore, the characterization vector contains n_A, n_G, n_T, and n_C.

2) The second numerical parameter is the sum of distances of each nucleotide base to the first nucleotide. Total distance Ti is defined as:

$$T_i = \sum_{j=1}^{n_i} t_j$$

i = A, G, T, C; t_j is the distance from the first nucleotide to the j^{th} nucleotide i in the DNA sequence.

3) The third parameter selected for the vector analysis is the distribution of each nucleotide along the DNA sequence. The variance of distance for each nucleic base used to describe the distribution is defined as the following:

$$D_i = \sum_{j=1}^{n_i} \frac{(t_j - \mu_i)^2}{n_i}$$

where i = A, T, G, C; tj is the distance from the first nucleotide to the jth nucleotide i in the DNA sequence and

$$\mu_i = \frac{T_i}{n_i}$$

So, the characterization vector, which contains twelve-dimensional information, is given as follow:

$$< n_{\mathrm{A}}, T_{\mathrm{A}}, D_{\mathrm{A}}, n_{\mathrm{G}}, T_{\mathrm{G}}, D_{\mathrm{G}}, n_{\mathrm{T}}, T_{\mathrm{T}}, D_{\mathrm{T}}, n_{\mathrm{C}}, T_{\mathrm{C}}, D_{\mathrm{C}} >$$

Conclusion

The proposed scalable kernelized fuzzy clustering algorithms based on in-memory computation for handling big data can be applied for clustering of DNA, and single nucleotide polymorphisms (SNP) sequences of soybean and other plant species .

Future Work

Experimentation will be performed on SNP and DNA datasets to show the effectiveness of proposed KSRSIO-FCM in comparison with KSLFCM.

References

[1] Cai, Weiling, Songcan Chen, and Daoqiang Zhang. "Robust fuzzy relational classifier incorporating the soft class labels." *Pattern Recognition Letters* 28.16 (2007): 2250-2263.

[2] Bharill, Neha, Aruna Tiwari, and Aayushi Malviya. "Fuzzy based scalable clustering algorithms for handling big data using apache spark." *IEEE Transactions on Big Data* 2.4 (2016): 339-352.

[3] Liu, Libin, Yee-kin Ho, and Stephen Yau. "Clustering DNA sequences by feature vectors." *Molecular phylogenetics and evolution* 41.1 (2006): 64-69.