Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186-5140
Volume 8, Number 1, pages 14–18, January 2019

# Performance Evaluation of Fuzzy C-means Associated with Spark based on Virtual Cloud Computing

Gui-Ren Lin
*National Formosa University*
*Department of Computer Science and Information Engineering*
Yunlin County 632, Taiwan
guir.lin5631@gmail.com

I-Ching Hsu
*National Formosa University*
*Department of Computer Science and Information Engineering*
Yunlin County 632, Taiwan
hsuic@nfu.edu.tw

***Abstract*—In recent years, machine learning and cloud computing technology have been widely studied and applied. Machine learning interprets potential knowledge by processing big data. Cloud computing can improve the efficiency of machine learning to process big data. Fuzzy C-Means (FCM) is a machine learning technique that is applied to data clustering. Apache Spark is Open Source cluster cloud computing framework that mainly applies in-memory computing technology. This study proposes an Integrating Fuzzy C-Mean and Spark based on Virtual Cloud Computing Framework (IFCSVCCF) that focuses on the use of Fuzzy C-Mean and Spark to handle big data based virtual architectures.**

***Keywords—machine learning, cloud computing, spark***

## I. Introduction

Machine learning builds models through training and testing to perform predictions[1]. Fuzzy C-Means (FCM) is a machine learning technique that is applied to data clustering. FCM is a method of clustering which assigns one object to several clusters so that one object can belong to multiple clusters. The concept of FCM algorithm is to calculate the point of the clustering center as well as the distance between the center and each object through object distribution. Then it will constantly modify the clustering center to get the optimal solution of clustering center value, apply Euclidian Distance to calculate the distance between each object and each clustering center, and then use these distances to calculate the probability for these objects belonging to each category to achieve fuzzy effect.

Apache Hadoop [2] is an open source and is a quite popular cloud computing platform. The core technologies of Hadoop consist of HDFS and Map Reduce. Apache Spark[3] is open source cluster cloud computing framework with the core of RDD (Resilient Distributed Dataset). It mainly applies in-memory computing technology which enables computing through the memory before the data is written into the hard disk. Theoretically, the Spark work up to 100 times faster than Hadoop, and could be 10 times faster than Hadoop even if Spark is executed simultaneously in the hard disk.

Docker is a lightweight class virtualization technology. Compared with traditional virtualization technology, it has such features as rapid allocation and low resource utilization. Traditional visual machine system manages each visual machine through Hypervisor and must set up operation system on each visual machine, while Docker manages with Docker Engine and each visual machine can be regarded as a container[11]and can directly set up required environment without the need for additional operation system.

This study proposes an Integrating Fuzzy C-Mean and Spark based on Virtual Cloud Computing Framework (IFCSVCCF) that focuses on the use of Fuzzy C-Mean and Spark to handle big data based Docker virtual cloud computing environment.

## II. Related work

Many medicines are used in the world nowadays. Upon the research on medication information. In [4], author integrated the ATC code (Anatomical Therapeutic Chemical code) formulated by World Health Organization and the NDF-RT (National Drug File Reference Terminology) of the United States with Semantic Web Technologies (SWT). In terms of the medication interaction, Abha Moitra et al[5] applied Semantic Web technology and Semantic Application Design Language (SADL) to set up the Answer Set Programming (ASP) for the concepts in Drug Interaction Knowledge Base (DIKB) to understand which medications may have interactions. Asma Ben Abacha et al[6] proposed integration of Support Vector Machine (SVM) and a kernel-based approach to predict the interactions between medications. Dries Harnie et al [7] executed computing on machine learning through Apache Spark and to find new medicines. Moreover, this study also compared the speed of different output file type and efficiency of different nodes.

## III. System Design

This study proposes an Integrating Fuzzy C-Mean and Spark based on Virtual Cloud Computing Framework (IFCSVCCF), as shown in Figure 1. Such framework consists

of five parts: IoT Module, Open Data Module, Machine Learning Module, Semantic Web Module and Cloud Virtualization Module, which are introduced as below:

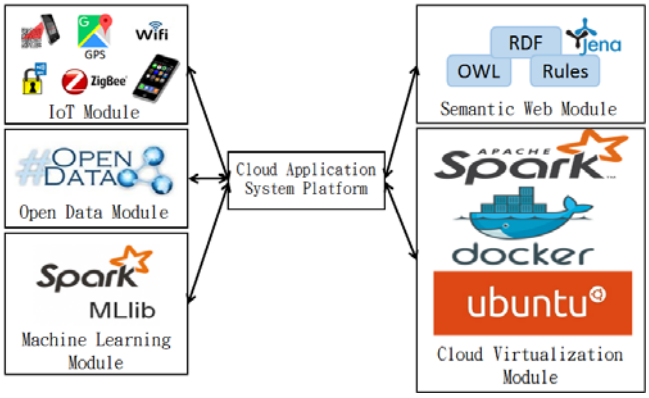


Fig. 1. The IFCSVCCF

### A. *IoT Module*

IoT module consists of the smart pillbox, mobile App and cloud management platform. When time is up, it can remind the senior patients to take pills and immediately inform the patient's family member of the patient's unusual medication status or forgetting to take pills.

### B. *Open Data Module*

The open data in Open Data Module is provided by Taiwan government's open data platform. The open data used in this study include: "healthcare medication item query document", "all drug license dataset", "drug appearance dataset", "basic pharmacy dataset" and "food ingredient dataset". The application of open data enhanced the authenticity of this study.

### C. *Machine Learning Module*

Machine Learning Module makes food categorization with Jieba and Fuzzy C-Means(FCM)algorithms. It first extracts related words with Jieba tokenization and then categorizes the foods with FCM algorithm to provide users with list of foods that are worth of their attention.

### D. *Semantic Web Module*

This study uses Semantic Web Module to connect the whole system. When the user imports health bank data, the system will automatically generate RDF file and make a series of inference. The inference results include medicines, foods, food ingredients that are worth attention as well as sensing of the drug weight.

### E. *Cloud Virtualization Module*

Cloud Virtualization Module uses Spark cluster cloud computation platform, which can connect multiple computers and disperse the data into each computer to make computing with certain error-tolerant rate in order to avoid possible failure caused by the huge data size. Besides, this module also applies Docker virtualization technology to cut the resources in one computer into several visual computers which can make computing simultaneously. The virtualization technology makes the cloud computing more flexible.

## IV. System Development

This study developed Safety Medication Cloud Platform (SMCP) to verify the feasibility of IFCSVCCF. The system integrate IoT(Internet of Things), Fuzzy C-means(FCM) algorithm, Semantic Web and Cloud Computing technology to provide senior patients and their relatives with more abundant personalized medication information and avoid the problem of forgetting to take pills. With the increase of food related open data and health bank data, the system also provides cloud computing technology to process the huge data. The SMCP architecture is shown in Figure 2 and the steps are illustrated as below.

1 User

1.1 The user imports the senior's health bank data onto the safety medication cloud platform.

1.2 The user can view the medication data via mobile App or webpage platform, including the senior patient's personalized medication information inquiry, drug inquiry, drug and food that must be avoided and nearby pharmacy inquiry functions.

2 Manager

2.1 The manager can import the open data onto the safety medication cloud platform. If the imported data is the food open data, the system will extract the eigenvector of each food with Jieba tokenization for the purpose of machine learning categorization.

2.2 Execute FCM algorithm on Spark. After the execution is completed, it will make post processing of the data and categorize the food in the open data into specific categories according to algorithm results.

3 Sensing data

3.1 The smart pillbox will transmit the sensed drug data onto the safety medication cloud platform. It can judge whether there is any case of forgetting to take pills or unusual medication.

4 Computing

4.1 Through Jena, it can infer from the user's medication data more matters worth attention, including drugs and food that must be avoided, and whether there is any error in the drug weight. SMCP will apply virtualization technology and cloud computing technology to improve the efficiency of machine learning the Semantic Web inference.

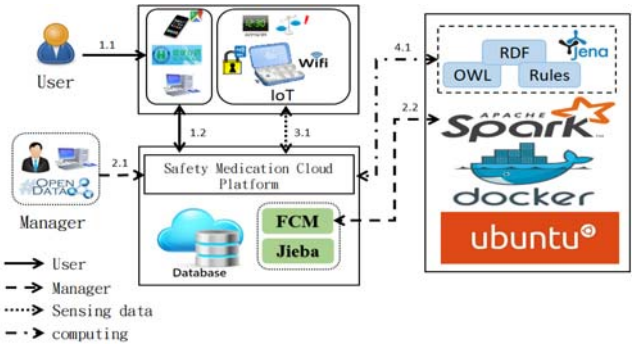


Fig. 2. The dataflow-oriented SMCP

## V. Performance Evaluation

This section provides a performance analysis of the proposed architecture and system.

### A. *Experimentation Environment*

#### 1) *Test environment 1*

Test environment 1 is single physical machine mode, which establish Spark on one computer, and make parallel computing through Spark on one machine. The CPU is I7-2600@3.4Gz, memory is 16GB and the HDD is 2TB. The operation system used for the test environment 1 is Ubuntu16.04 LTS.

#### 2) *Test environment 2*

Test environment 2 is physical machine cluster mode, which will establish 7 server hosts. The CPU for each machine is I7-2600@3.4Gz, the memory is 16GB and the HDD is 2TB. The operation system used for test environment 2 is Ubuntu16.04 LTS.

The main architecture of cluster test environment is on 7 computers. The cluster architecture used for this study is shown in Figure 3. It uses one master for cluster management and resource deploy. The Cluster Manager is the resource managers responsible for deploying the resources. Standalone resource manager is built in Spark and the developer can also choose other resource managers according to requirements. This study uses two resource managers, namely the Spark Standalone resource manager and Hadoop YARN resource manager to probe into the execution efficiency of Spark. The architecture and specifications of these two resource managers are introduced as below:

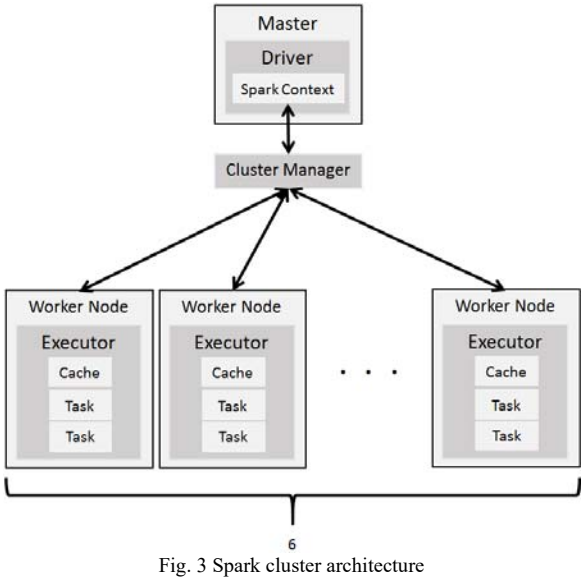


Fig. 3 Spark cluster architecture

#### 3) *Test environment 3*

Test environment 3 is virtual machine cluster mode. In this study, the computers used for irtual machine cluster test are the 7 ones used for physical machine cluster test. visual environment is constructed on each computer with Docker.

Docker achieves the virtualization with containers. Each container is an isolated platform. As the containers will utilize the resources of the native system flexibly, in the Docker it is not necessary to deploy additional resources to the containers. However, to avoid system shutdown caused by the container's depletion of resources, this study will limit the maximum resources that can be used by each container. This study uses 4-core 8-execution CPU I7-2600@3.4Gz. To avoid computer shutdown caused by insufficient resources, this study will set limitations based on the quantity of CPU cores. That is, the one core can be used by each container, and the remaining cost is kept for the host. Accordingly, the memory is divided into 4 shares, each with 4G, which will be allocated to each container and the host.

Figure 4 is the Docker VM cluster architecture. This architecture is composed by 7 physical machines with 3 containers created by Docker on each physical computer. In this environment, one Container will serve as the Master to deploy tasks, while the other 20 Containers will be the Slaves. Spark and Hadoop environments will be established in all Containers. The purpose of establishing Hadoop environment is to allow Spark to use the Hadoop YARN resource dispenser and HDFS file system. After the establishment, it can test the VM cluster.

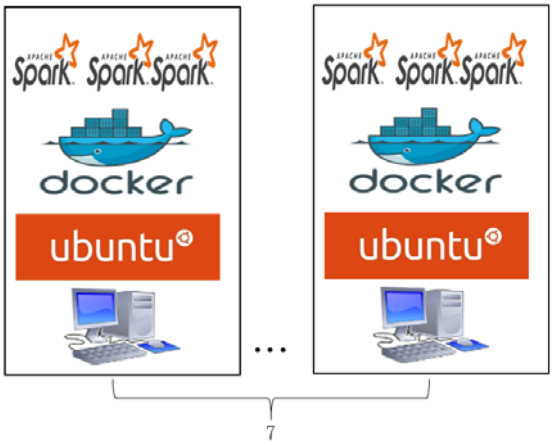


Fig. 4.VM cluster framework

## B. *Cloud computing cluster mode*

### 1) *Standalone*

This mode will use Spark Standalone as the resource manager, the stacking architecture of which is shown in Figure 5. Standalone provides simple scheduling function and fault-tolerant mechanism and is easy to use and establish. As this mode uses HDFS file system, it must install both Spark and Hadoop platform environments on each computer and start HDFS file system.

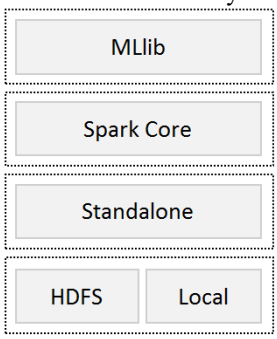


Fig. 5. Spark Standalone

### 2) *YARN*

This mode uses YARN as the resource manager with the stacking architecture shown in Figure 6. YARN provides multiple scheduling functions and fairly good fault-tolerance mechanism. This mode must install both Spark and Hadoop platform environments on each computer and start HDFS file system and YARN cluster resource manager.

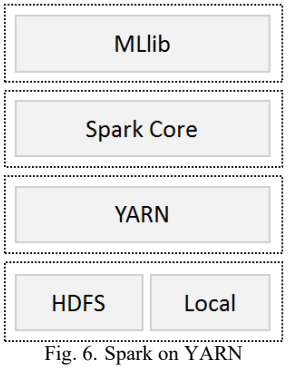


Fig. 6. Spark on YARN

## C. *Big data test and result*

Big data test mainly tests the differences among Test2 Standalone mode, Test3 Standalone mode, Test2 YARN mode and Test3 YARN mode when the data size is large.

Big data test result is shown as Figure 7. Test 2 Standalone mode, Test 3 Standalone mode and Test 3 YARN mode are similar in efficiency, while the efficiency of Test 2 YARN mode is much lower than the above three. The reason is that the advantages of YARN can only reveal when there are many node points, so when the machines are the same, if the number of node points can be increased with virtualization software, the efficiency would be much between than using physical ones. Although Standalone will have similar efficiency no matter it is visualized or not, its stability and resource scheduling effect is not as good as YARN. So, Standalone can be applied during development or test stage, as it can accelerate the development progress with its advantages of easy installation and fast speed. When there are many node points or under actual implementation situation, it is suggested to use YARN as it can hoist the instruction cycle and make the system more stable.

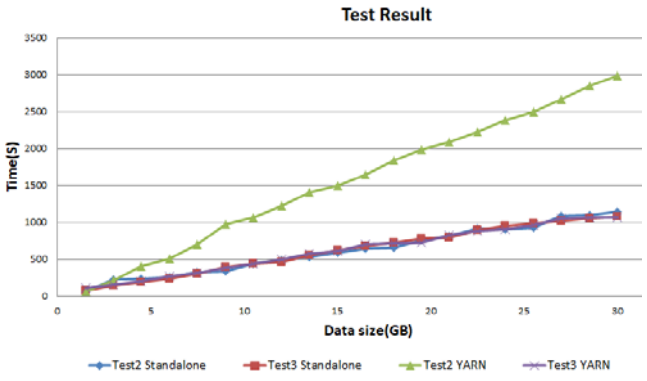


Fig. 7. Big data test result

## VI. Conclusion

This study proposes an Integrating Fuzzy C-Mean and Spark based on Virtual Cloud Computing Framework (IFCSVCCF) that focuses on the use of Fuzzy C-Mean and Spark to handle big data based Docker virtual cloud computing environment. We design three different Apache Spark test environment associated with two different manager model. The

results of the preliminary tests show that using a virtual cloud computing architecture will have better performance in our proposed the IFCSVCCF and test environment. In the future, we will integrate different machine learning technologies, such as SVM, K-means[8], K-NN[9], etc., to test our proposed architecture.

## ACKNOWLEDGMENT

The author would like to thank the Ministry of Science and Technology, R.O.C, for financially supporting this research under Contract No. MOST 107-2637-E-150-009-

## REFERENCES

[1] Q. Zhang, L. T. Yang, Z. Chenc, and P. Lic, "A survey on deep learning for big data," Information Fusion, vol. 42, pp. 146-157, 2018.

[2] Apache. (2017, Aug. 30, 2018). Hadoop-Apache. Available: http://hadoop.apache.org/

[3] J. Domann and A. Lommatzsch, "A highly available real-time news recommender based on Apache Spark " Lecture Notes in Computer Science vol. 10456 pp. 161-172, 2017.

[4] Q.Zhu, "Pharmacological Class Data Representation in the Web Ontology Language (OWL)," 2014, pp. 77-84.

[5] A.Moitra, R.Palla, and L.Tari, "Semantic Inference for Pharmacokinetic Drug-Drug Interactions," in IEEE International Conference on Semantic Computing Newport Beach, CA, USA.

[6] A.Ben, F.Mahbub, A.Karanasiou, Y.Mrabet, A.Lavelli, and P.Zweigenbaum, "Text mining for pharmacovigilance : Using machine learning for drug name recognition and drug – drug interaction extraction and classification," J. Biomed. Inform., vol. 58, pp. 122–132, 2015.

[7] D. Harnie, A. E. Vapirev, and J. K. Wegner, "Scaling Machine Learning for Target Prediction in Drug Discovery using Apache Spark," in 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing Shenzhen, China

[8] R. K. Roul and S. K. Sahay, "K-means and wordnet based feature selection combined with extreme learning machines for text classification," presented at the 12th International Conference on Distributed Computing and Internet Technology, ICDCIT 2016 Bhubaneswar, India, 2016.

[9] M. U. Bokhari, Q. M. Shallal, and Y. K. Tamandani, "Reducing the Required Time and Power for Data Encryption and Decryption Using K-NN Machine Learning," IETE Journal of Research, pp. 1-9, 2018.