Journal of Ubiquitous Systems & Pervasive Networks
Volume 3, No. 1 (2011) pp. 00-00

An Enhanced Fuzzy Clustering to Pattern Recognition for Cloud Computing, by using Model Aggregation and Model Selection

Choukri Djellali *, Mehdi adda , Mohamed Tarik Moutacalli

Department of Computer Science and Engineering
University of Quebec At Rimouski
Rimouski, Canada, G5L 3A1

Abstract

Numerical schemes research on clustering models has been quite intensive in the past decade. Many models have been proposed to address the clustering tasks. Most clustering models are influenced by presentation order, complex shapes, architecture configuration, and learning instability. Hence, in the present study, a novel clustering-based method for cloud computing that provides an improvement in recognition rate, is described. The evaluation, based on 10-fold Cross-validation, showed that the proposed model, which is named BaggingCluster, yielded good results and performed better than Self Organizing Map and fuzzy Adaptive Resonance Theory. Experimental studies demonstrate that our model provides an efficient model for cloud computing.

Keywords: *Data Mining, Deep Learning, Clustering, Artificial Neural Network, Fuzzy Logic, Cross-Validation.*

1. Introduction

In Data Mining and pattern recognition, clustering is the task of assigning each input pattern to one of a set of clusters. It is considered as a separate class of unsupervised learning that analyzes the training patterns and produces the relevant model.
Clustering has been widely used in the Data Mining community, significantly improving the state-of-the-art models. It is one of the powerful Data Mining techniques that uses unsupervised learning in which the patterns are grouped on the basis of their similarities or mutual distances. It is defined as follows: Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters [1].
Clustering has been applied in a wide variety of fields, ranging from Cloud Computing [2], Internet of Everything (IoE)[3], Cognitive Computing [4], Big Data [5], Bioinformatics [6], and many other domains.

Most clustering algorithms are sensitive to noise, Bellman's curse of dimensionality, architecture configuration, and instability recognition.

On one hand, model aggregation is a powerful technique that combines multiple clustering algorithms to improve stability and accuracy.

On the other hand, Cross-Validation is a data-based machine learning technique applied to address bias-variance tradeoffs, and thus, to improve recognition accuracy.
Based on these premises, in the present study, we propose a new conceptual model to enhance the clustering accuracy based on model aggregation and model selection techniques.

This paper is divided into six sections in addition to the introduction. In Section 2, we present a brief overview of related work. The conceptual architecture of our approach is given in Section 3. Before we conclude, we give in Section 4 a short evaluation with benchmarking models for our conceptual model. Then, a conclusion (Section 5) ends the paper with future works (Section 6).

*Corresponding author. Tel.: +1-418-833-8800
Fax: +1-418-724-1525; E-mail: Choukri_Djellali@uqar.ca

DOI: 10.5383/JUSPN.03.01.000

2. Literature review

Several studies of knowledge clustering focused on the identification of relevant models to provide a rich understanding of a particular domain. It is the task of automatically sorting a set of patterns into clusters from a training set. This technique has been used successfully in several applications such as topic identifications, document organization, target marketing, etc. In this context, a wide variety of clustering models have been proposed in machine learning literature. These models include:

- Partitional Clustering: used to cluster examples within a data set into different clusters based on their similarity. Among the most popular models, we quote: K-means, K-medoids, CLARA, fuzzy C-Mean, Bisecting K-means, CLARANS, CLASA [7], etc.
- Hierarchical Clustering: is one of the most popular clustering methods, which is based on successive groupings that produces a binary classification tree also known as a dendrogram. We quote in this category the following models: Single link, Complete link, Group average, BRICH, CURE [7], etc.
- Density-based clustering : this model represents each cluster by a region of high point density. Among the most popular models in this category, we quote: DBSCAN, OPTICS, DECODE, DENCLUE [7], etc.
- Grid-based clustering: this model uses a grid structure of cells to represent the clusters. We quote in this category the following models: STING, OptiGrid, GRIDCLUS, WaveCluster [8], etc.
- Genetic clustering: involves combinatorial optimization process based on evolutionary computation in order to accelerate the convergence speed of training. Crossover and mutation operators are applied to provide high genetic diversity. We find in this category the following models: genetic-TS, Genetically guided algorithm or GGA, genetic k-medoid, simulated annealing fuzzy c-means or SA-FCM [8] and [7], etc .
- Artificial Neural Networks (ANNs) (also known as connectionist computational models): designed to simulate the biological Neural Networks. The neural architecture is composed of many interconnected units usually known as artificial neurons. It is widely used for more complex tasks such as categorization, prediction, clustering, regression, and summarization, etc. Among the most popular models in this category, we quote: Self Organizing Map (SOM), Growing Neural Gas (GNG), Adaptive Resonance Theory (ART), Real-Time Recurrent Learning (RTRL), Gated Recurrent Units (GRUs), Boltzmann Machine, Learning Vector Quantization (LVQ), Deep Belief Net- works (DBNs), Hopfield, Bidirectional Associative Memory (BAM), Growing Cell Structures (GCS), Recurrent Convolutional Neural Network (RCNN) [8], [9].

A more detailed study of different models can be found in the papers [8] and [7].

Machine learning has been significantly advanced through the use of clustering techniques. Many of the recent advancements have led to the appearance of several approaches for pattern recognition.

Tan et al. (2019) [10] introduced an application of Self-Organizing Feature Map Neural Network based on K-means clustering in Network Intrusion Detection. The data set used in this study comes from the NSL-KDD network intrusion detection database. Experimental results showed that this clustering model improved pattern recognition and significantly reduced the training time.

Riese et al. (2020) [11] presented a Supervised Self-organizing Maps (SuSi) framework, which used unsupervised, supervised, and semi-supervised classification on high-dimensional data. The evaluation based on soil moisture data sets showed that this Deep Learning model yielded good results and performed better than the random forest in the regression of soil moisture.

Nasser et al. (2019) [12] proposed a Deep Learning model that can be used for Predicting Movies Rates Category. The clustering architecture used an Artificial Neural Network for prediction. This model yielded good results and showed that it is able to 92.19% accurately predict the category of movies rate.

Elliot et al. (2020) [13] introduced a Deep Learning model for Cloud Computing Security in Banking Sector. The training step is based on the Levenberg Marquardt algorithm. The cuckoo search algorithm is used to improve the convergence speed of training. This Deep Learning model yielded good results with a False Rejection Rate of 0% and False Acceptance Rate of 8%.

Zhang et al. (2020) [14] proposed a method to identify the type and state of electric appliances based on a power time series. A Convolutional Neural Network was trained to identify the type of appliance. To compute the number of states of the appliance, a k-means algorithm was applied. The results showed that this model played a significant role in improving the accuracy of identifying both the type and the running state of electric appliances.

Recently, Nguyen Nasser et al. (2020) [15] proposed a Deep Learning model namely HKM–ANN, for predicting Blast-Induced Ground Vibration in an Open-Pit Mine, by using a Hybrid Model Based on Clustering and Artificial Neural Network. A Hierarchical K-Means clustering algorithm (HKM) was applied before training the predictive models. The proposed HKM–ANN model yielded good results compared to the state-of-the-art models.

Most pattern recognition models based on clustering are perturbed by noisy features, architecture configuration, and learning instability. Moreover, the decision boundaries yielded are not well separated.

Hence, in the present study, we propose an alternative numerical scheme to enhance the clustering results based on model aggregation and model selection.

3. Architecture of our clustering model

In this section, we introduce the architecture adopted in our approach. The clustering model learns from the training data and builds a relevant model. The process of clustering starts with presenting a set of patterns from the available examples in the data set as shown in Figure 1. The goal of training is to find the relevant model that captures the underlying structures of the pattern content. The uncovering of hidden structures is performed by model aggregation and model selection.

In order to categorize patterns according to their contents, we used a Bootstrap aggregation scheme based on fuzzy k-medoids and fuzzy C-Means [8], [7],

This meta-modeling technique improves the stability and accuracy of clustering algorithms.

In order to minimize the variance and bias of our model, we used an effective sampling technique based on k-fold Cross-Validation. The data set is sampled into a training set and testing

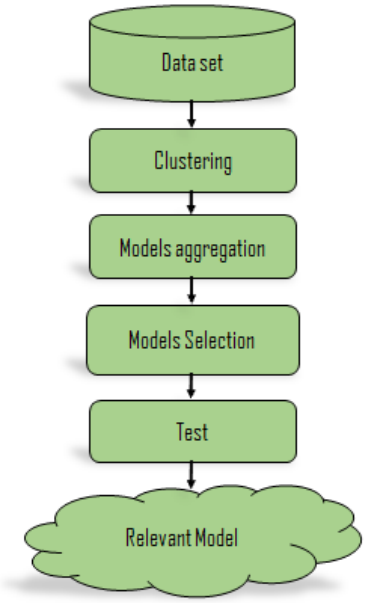

Fig. 1. Clustering model.

set. Of the k blocks, a single block is retained as the test data for performance evaluation, and the remaining (k-1) blocks are used as training samples. The aggregate model created from a combination of aggregated models improves the stability and accuracy. Finally, the test sets are used for evaluating the performance of the techniques used for pattern recognition.

The next section explores the baselines and benchmark Metrics in more detail and describes the results and discussions of experiments conducted

4. Experimental study

Through experimental studies, we first present the used data set for training. We describe the benchmarking models and measures used for performance evaluation, and demonstrate the ability of our model to reach the optimal solution.

4.1. Configuration

Our proposed architecture has been implemented on Neon.1a Release (4.6.1) eclipse integrated development environment 64-bit and some library functions such as JDK 11.0.6 + Java EE, Java Matrix Package or JAMA[1], etc.

4.2. Data set

In our study, we used the QoS Data set for cloud computing[2], which is the widely used data set for cloud computing. The data were divided into 70% for training and the remaining 30% for testing. Each input vector contains 5825 features in which each feature represents the time duration between request sending and response receiving. The average response time is equal to 0.8111 ms and the standard deviation is equal to 1.9670 ms.
As shown in Figure 2, from the 339 service users, we select the first three service users and we plot their response time value. The X-axis represents the Web service and Y-axis shows the response time value.

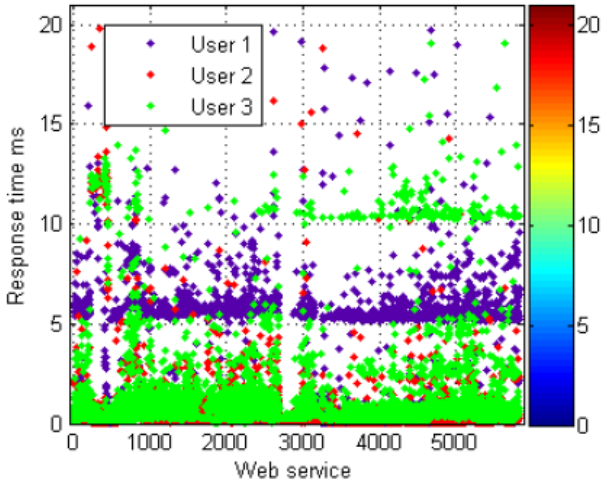

Fig. 2. Data set.

4.3. Model aggregation

As previously mentioned, two clustering models are used in our approach as an aggregated models.

- Fuzzy C-Means or FCM: is a soft clustering that assigns a data of n input vectors $X = \{x_1, x_2, ..x_n\}$ into a set of c fuzzy cluster $C = \{\omega_1, \omega_2, .., \omega_c\}$ according to the degree to which an input vector x_i belongs to cluster ω_j. Each cluster indicates the strength of the association u_{ij} between the input vector x_i and a particular centroid ω_j.
This fuzzy clustering algorithm is based on minimization of the following loss function:

$$L_\pi = \arg\min_c = \sum_{i=1}^{i=n} \sum_{j=1}^{c} u_{ij}^m \|x_i - \omega_j\| \quad (1)$$

The iterative optimization updates the membership u_{ij}^m and centroids ω_j by the following formulas:

$$u_{ij}^m = \frac{1}{\sum_{k=1}^{k=c} \left(\frac{\|x_i - \omega_j\|}{\|x_i - \omega_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

$$\omega_j^m = \frac{\sum_{i=1}^{i=n} u_{ij}^m x_i}{\sum_{i=1}^{i=n} u_{ij}^m} \quad (3)$$

Where u_{ij} stands for the degree of membership and c is for the number of clusters.

[1] http://math.nist.gov/javanumerics/jama/

[2] https://github.com/wsdream/wsdream-dataset

- fuzzy k-medoids: is an efficient partitional clustering algorithm used for cluster analysis in Data Mining and Machine Learning. Each medoid minimizes the average distance to all patterns in the cluster. Hence, the optimal number of medoids should minimize a loss function as follow:

$$\begin{cases} O_{\pi^*} = \arg\min_{W,Z} = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ji}^m d(v_i, m_j) \\ 0 \leq u_{ji} \leq 1, \ \forall ij, \ 0 \leq i \leq n, \ 0 \leq j \leq k \\ \sum_{j=1}^{k} u_{ji} = 1, \forall i, \quad 1 \leq i \leq n \\ 0 < \sum_{i=1}^{n} u_{ji} < n, \forall j, \ 1 \leq j \leq k \end{cases} \tag{4}$$

Where W is an $n \times k$ fuzzy matrix containing the membership degree, n is the number of patterns in the data set, k is the number of clusters, $Q = \{m_j\}_{j=1}^{k}$ is a set of medoids.
The fuzzy k-medoids algorithm updates the membership degree u_{ij} by the following formula:

$$u_{ij}^m = \frac{(\frac{1}{d(v_i, m_j)})^{\frac{1}{m-1}}}{\sum_{i=1}^{k}\left(\frac{1}{d(v_i, m_i)}\right)^{\frac{1}{m-1}}} \tag{5}$$

We used Mean Square Error (MSE) as a measure of how well the models fit data, which is the average squared difference between the desired outputs and current outputs.

$$MSE = \frac{1}{n_L}\sum_{j=1}^{n_L}(O_j^C - O_j^D)^2 \tag{6}$$

Where O_j^C stands for current output, O_j^D is for desired output, and n_L is for the number of clusters.

Figure (3) shows the learning curve as a function of the amount of training error. The X-axis indicates the number of epochs or presentations of the full training set and Y-axis shows the amount of error. The training error does not decrease monotonically, it generally decreases, it can increase or oscillate.
At the start of the learning, the curve shows a high error which indicates that the training input patterns are not well separated. After training, the curve shows a low error and this means the learned model tends to be close to the training patterns of the data set. Hence, our clustering model searches the complex shape of the decision boundaries and avoids local minima during training. It maximizes the generalization ability and induces effectively the relevant model.
The estimate generalization is not based on a cost function but the error of our trained clustering model. The recognition accuracy of fuzzy k-medoids is equal to 97.11% after 79 iterations. Compared to fuzzy k-medoids, the recognition accuracy of fuzzy C-Means is equal to 96.77% with more training, i.e., the number of complete passes through the training patterns is equal to 101.

In the next section, we review the benchmarking models and performance measures used to test the effectiveness of our clustering model.

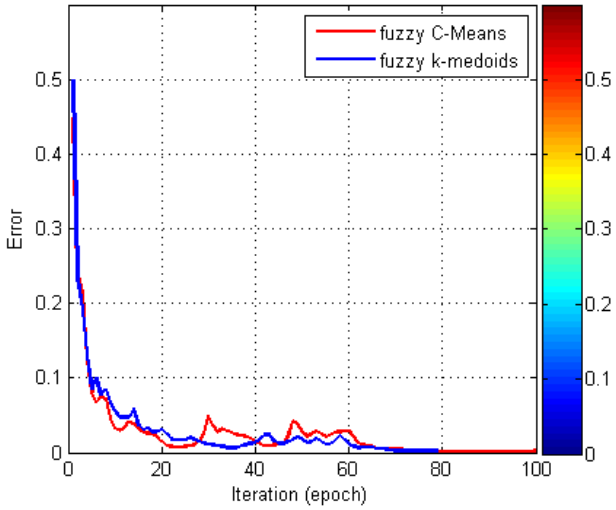

Fig. 3. Learning Error V.S Iteration

4.4. Evaluation

Our Experimental study was designed to compare three clustering models namely Self Organizing Map, fuzzy Adaptive Resonance Theory, and our aggregated model named BaggingCluster.

- Fuzzy Adaptive Resonance Theory: this Neural Network is governed by two subsystems. The attention subsystem provides a winning neuron (or cluster) and the orientation subsystem decides the resonance acceptance. They are interacting with bottom-up and top-down processes of long-term memory or LTM. The neural network is in the state of resonance if the orientation system accepts a winning neuron, i.e., when the prototype neuron accepts a unification with the current input pattern according to a resonance threshold. If the resonance does not appear, the orientation system allows the attention subsystem to increase its resources to meet the external requirements, i.e., dynamic online learning [7].
- Self Organizing Map (also known as topology preserving maps): is an Artificial Neural Network that contains an input layer and output layer. Unsupervised training based on a competitive mechanism is used to modify the connection weights between neurons [8].

In order to obtain stable scoring results, we applied a model selection technique based on 10-fold Cross-Validation. By running repeated 10-fold Cross-Validation on training patterns, the aggregate estimation is defined as the average of the estimations obtained on each fold.

$Precision$, $Recall$ and $F-measure$ measures are used for performance evaluation, which are calculated for 10 folds.

$$Precision_\mu = P = \frac{\sum_{i=1}^{c} tp_i}{\sum_{i=1}^{c} tp_i + fp_i} \tag{7}$$

$$Recall_\mu = R = \frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k} tp_i + fn_i} \tag{8}$$

$F-measure_\mu$ index weights average of the $precision_\mu$ and $recall_\mu$, i.e.,

$$F - measure_{\mu} = 2 \times \frac{precision_{\mu} \times recall_{\mu}}{precision_{\mu} + recall_{\mu}} \quad (9)$$

where tp, fp and fn are true positive, false positive, and false negative, respectively.
The classification accuracy of clustering models and other baselines are shown in Table 1.

Table 1. The effectiveness of pattern recognition.

Model	P	R	$F - measure$	Time
SOM	82.17	77.89	79.97	3077
$fuzzy\ ART$	83.17	82.13	82.65	3107
$BaggingCluster$	87.15	82.77	84.90	6199

Experimental results show that our clustering model has good performance, which provides an efficient model for cloud computing. However, the processing time for SOM is equal to 3077 microseconds and 3107 microseconds during $fuzzy\ ART$. Compared to SOM and Fuzzy ART, the processing time of $BaggingCluster$ is equal to 6199 microseconds.
One of the main advantages of our clustering model is its ability to directly construct the complex decision boundaries around patterns; therefore, the lowest average generalization error of clustering.

5. Conclusion

In this paper, we have introduced a clustering model providing performance analysis to pattern recognition for cloud computing. Our approach exploits model aggregation, and model selection techniques to increase the productivity of knowledge extraction. We used the model aggregation technique to find a single consolidated clustering model for better data fitting. The model selection technique is applied to optimize the bias-variance tradeoff of the expected prediction of our clustering model.
Hence, the Bootstrapping scheme based on 10-fold Cross-Validation and model aggregation is indispensable to improve generalization ability.
We demonstrate its ability to reach the optimum solution and obtain more cluttered decision boundaries. Experiments show that our model has good performance, which provides an efficient clustering model for cloud computing.
Our next work is on Deep Learning where the goal is to produce an enhanced generalization by using boosting, ensemble learning and model selection techniques.

Acknowledgments

We would like to gratefully acknowledge the support of the University of Quebec at Rimouski for funding this research.

References

[1] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[2] Muhammad Shafie Abd Latiff, Syed Hamid Hussain Madni, Mohammed Abdullahi, et al. Fault tolerance aware scheduling technique for cloud computing environment using dynamic clustering algorithm. *Neural Computing and Applications*, 29(1):279–293, 2018.

[3] Anam Javaid, Asma Rafiq, Maaz Rehan, M Mustafa Rafique, M Kamran, and Ehsan Ullah Munir. Clustering-cum-handover management scheme for improved internet access in high-density mobile wireless environments. *Sustainable Computing: Informatics and Systems*, page 100483, 2020.

[4] Kai Hwang and Min Chen. *Big-data analytics for cloud, IoT and cognitive computing*. John Wiley & Sons, 2017.

[5] Farouk Ouatik, Mohammed Erritali, and Mostafa Jourhmane. Student orientation using machine learning under mapreduce with hadoop. *J. Ubiquitous Syst. Pervasive Networks*, 13(1):21–26, 2020.

[6] Quan Zou, Gang Lin, Xingpeng Jiang, Xiangrong Liu, and Xiangxiang Zeng. Sequence clustering in bioinformatics: an empirical study. *Briefings in bioinformatics*, 21(1):1–10, 2020.

[7] Peter E Hart, David G Stork, and Richard O Duda. Pattern classification. *John Willey & Sons*, 10, 2001.

[8] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2): 165–193, 2015.

[9] Christophe Feltus. Ai's contribution to ubiquitous systems and pervasive networks security-reinforcement learning vs recurrent networks. *J. Ubiquitous Syst. Pervasive Networks*, 15(02):1–9, 2021.

[10] Ling Tan, Chong Li, Jingming Xia, Jun Cao, et al. Application of self-organizing feature map neural network based on k-means clustering in network intrusion detection. *Computers materials & Continua*, 61(1):275–288, 2019.

[11] Felix M Riese, Sina Keller, and Stefan Hinz. Supervised and semi-supervised self-organizing maps for regression and classification focusing on hyperspectral data. *Remote Sensing*, 12(1):7, 2020.

[12] Ibrahim M Nasser, Mohammed O Al-Shawwa, and Samy S Abu-Naser. A proposed artificial neural network for predicting movies rates category. 2019.

[13] SJ Elliot, VIE Anireh, and ND Nwiabu. A predictive model for cloud computing security in banking sector using levenberg marquardt back propagation with cuckoo search. 2020.

[14] Ying Zhang, Bo Yin, Yanping Cong, and Zehua Du. Multi-state household appliance identification based on convolutional neural networks and clustering. *Energies*, 13 (4):792, 2020.

[15] Hoang Nguyen, Carsten Drebenstedt, Xuan-Nam Bui, and Dieu Tien Bui. Prediction of blast-induced ground vibration in an open-pit mine by a novel hybrid model based on clustering and artificial neural network. *Natural Resources Research*, 29(2):691–709, 2020.