Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

Artículo original

Using K-means algorithm for regression curve in big data system for business environment

Usando el algoritmo K-means para la curva de regresión en un gran sistema de datos para el entorno empresarial

Mohammed Anouar Naoui [1]: https://orcid.org/0000-0003-1653-531X
Brahim Lejdel[2] : https://orcid.org/0000-0003-1779-0689
Mouloud Ayad[3 *]: https://orcid.org/0000-0001-9858-8612

[1]LIMPAF Laboratory, Computer science department,Faculty of Sciences and applied Sciences. University of Bouira.
[2] Computer science department, University of El-Oued.
[3] LPM3E Laboratory, Faculty of Sciences and applied Sciences, University of Bouira.

[*] Corresponding Author: manouarn@yahoo.com

ABSTRACT

Predictive analysis quickly becomes a decisive advantage for desired range of Business activities. It involves methods and technologies for organizations to identify models or patterns for data. Big data bring enormous benefits to the business process. Big data properties such as volume, velocity, variety, variation and veracity, render the existing techniques of data analysis not

sufficient. Big data analysis requires the fusion of regression techniques for data mining with those of machine learning. Big data regression is an important field for many researchers, several aspects, methods, and techniques proposed. In this context, we suggest regression curve models for big data system. Our proposition is based on cooperative MapReduce architecture. We offer Map and Reduce algorithms for curve regression, in the Map phase; data transform in the linear model, in the reduce phase we propose a k-means algorithm for clustering the results of Map phase. K-means algorithm is one of the most popular partition clustering algorithms; it is simple, statistical and considerably scalable. Also, it has linear asymptotic running time concerning any variable of the problem. This approach combines the advantage of regression and clustering methods in big data. The regression method extract mathematic models, and in clustering, k-means algorithm select the best mathematic model as clusters.

Keywords: Cooperation MapReduce algorithm; Big Data; Regression Curve; k-means algorithm; Business environmental scanning.

RESUMEN

El análisis predictivo se convierte rápidamente en una ventaja decisiva para la gama de actividades comerciales deseadas. Implica métodos y tecnologías para que las organizaciones identifiquen modelos o patrones de datos. Los grandes datos aportan enormes beneficios al proceso empresarial. Las grandes propiedades de los datos, como el volumen, la velocidad, la variedad, la variación y la veracidad, hacen que las técnicas existentes de análisis de datos no sean suficientes. El análisis de grandes datos requiere la fusión de las técnicas de regresión para la minería de datos con las de aprendizaje automático. La regresión de grandes datos es un campo importante para muchos investigadores, varios aspectos, métodos y técnicas propuestas. En este contexto, sugerimos modelos de curvas de regresión para grandes sistemas de datos. Nuestra propuesta se basa en la arquitectura cooperativa de MapReduce. Ofrecemos algoritmos Map y Reduce para la regresión de la curva, en la fase Map; la transformación de datos en el

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

modelo lineal, en la fase reduce proponemos un algoritmo k-means para agrupar los resultados de la fase Map. El algoritmo K-means es uno de los algoritmos de clustering de particiones más populares; es simple, estadístico y considerablemente escalable. Además, tiene un tiempo de ejecución asintótica lineal en relación con cualquier variable del problema. Este enfoque combina la ventaja de los métodos de regresión y agrupación en grandes datos. El método de regresión extrae modelos matemáticos, y en la agrupación, el algoritmo k-means selecciona el mejor modelo matemático como agrupaciones.

Palabras clave: Algoritmo de cooperación MapReduce; Big Data; Curva de Regresión; algoritmo k-means; exploración del entorno empresarial.

Recibido: 16/12/2019

Aceptado: 31/03/2020

INTRODUCCIÓN

Regression analysis (Golberg et al., 2004) is a statistical methodology describe the relationship between variables atributes. For example in business marking, regression analysis can explain the relation between price and quality of products. The potential sales of a new product given its price. Regression analysis most used in continuous valued. Linear Regression (Bollobás.,1990) is a model describe the relationship between variables by linear model, let y is the response variable, and x the predictor variable, the model is:

$$y = ax + b \quad (1)$$

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

Where a and b can be solved by the method of least squares. Which minimize the error and extract the best line equation. if D set of data.D = {(x1, y1), (x1, y1), ...(xi, yi), ...(xn, yn)}

$$b = \bar{y} - \bar{x} \quad (2)$$

Where a and b can be solved by the method of least squares. Which minimize the error and extract the best line equation. if D set of data . D = {(x1, y1), (x1, y1), ...(xi, yi), ...(xn, yn)}.

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

Multiple linear regression

Relation between more than one variable describe by linear model, the general equation is:

$$y = a_1x_1 + a_2x_2 + ...a_lx_l + b \quad (4)$$

Non Linear Regression: Curve regression

Often the relationship between variables is far to being linear. Curve models are the most used, to determine the curve model relationship, there are several mathematics models such as power, exponential, logistic and polynomial model. We are going to present, in the Table 1, the multiple Curve models.

Table 1 - Curve regression models.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

Model	General expression
Power Model	$y = bx^a$
Exponential Model	$y = be^a x$
logistic model	$y = l - e^{(ax+b)}$
Polynomial model	$y = a_n x^n + a_{n-1}x^{n-1} + ...a_1 x^1 + b$

Once we have chosen the model to adopt, we must transform the curve into a Linear relation. There are several linearization methods which can be cited in Table 2:

Table 2 - Linearization Curve regression models.

General expression	linearization	$y' =$	$x' =$
$y = bx^a$	$log(y) = alog(x) + log(b)$	$y' = log(y)$	$x' = log(x)$
$y = be^{ax}$	$ln(y) = ax + ln(y)$	$y' = ln(y)$	$x' = x$
$y = l - e^{(ax+b)}$	$ln(y) = ax + b - ln(l)$	$y' = ln(y)$	$x' = x$
$y = a_n x^n + a_{n-1}x^{n-1} + ...a_1 x^1 + b$	$y = a_1 x_1' + a_2 x_2' = +...a_n x_n' + b$	$x_n' = x^n, x_2' = x^2, x_1' = x^1$	

Big data MapReduce Algorithms

MapReduce (Dean et al.,2010) primitives implements parallel processing, it composes by two algorithms, Map and Reduce, Map algorithm takes a set of data and convert it into another set of data. It takes a pair of (key, pair) and emits (key, pair) into Reduce algorithm. The input of Reduce algorithm is the result of map algorithm. The Map reduce constitutes from Master called Jobtracker, and a set of slaves server called TaskTracker (Shafer et al.,2010; Martha et al.,2013). Hadoop (Krishna.,2010) provide MapReduce runtimes with fault tolerance and dynamic flexibility support.

The essential question of our work are:

- What is the model that can present regression curve in big data system

This paper is organized as follows, in section 2. We present related works, linear model, curve regression and k-means algorithm. In section 3., we present our proposition, mathematic model, Map and Reduce algorithms and workflow architecture. Subsequently, we show in section 4.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

Validation and results of our proposition of UnversalBank data set. Finally, we terminate by the conclusion in section 6.

Related work

There are several research interested by regression, linear or curve in big data (Jun et al.,2015; Oancea et al.,2015; Ma et al., 2015; Neyshabouri et al., 2016). Several works oriented to propose mathematic approaches for regression in big data such as data (Jun et al.,2015; Ma et al., 2015; Neyshabouri et al., 2016). Other geared to proposes MapReduce algorithms and its implementations in big data system like (Oancea et al.,2015)

Linear model

(Jun et al. 2015) presented a divided regression analysis using multiple linear where regression form is :

$$y = a_1 x_1 + a_2 x_2 + ...a_m x_m + b$$

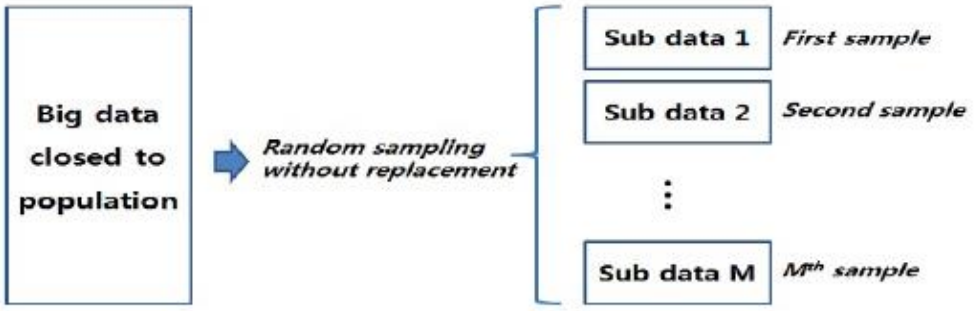


Fig. 1 - Dividing Big Data using Sampling Method (Jun et al.,2015).

Authors use random sampling data to divided big data into sub samples, they consider all attributes have an equal chance to be selected in the sample Figure 1. (Oancea et al. 2015)

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

presents a way to solve linear regression in big data, they propose a MapReduce algorithm expressed to the least square error, for the implementation they use R-Studio and Rhadoop library. (Ma et al. 2015) presented Leveraging for big data regression.Leverage appear, If a data point A is moved up or down, the corresponding adjusted value moves proportionally. The proportionality constant is called the leverage effect.

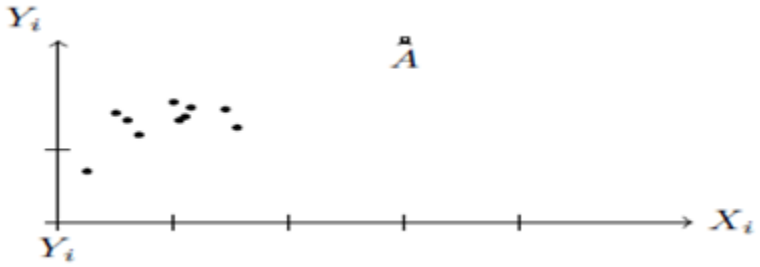


Fig. 2 - Example of leverage point A.

They propose two algorithms, Weighted Leveraging and Unweighted Leveraging algorithms for linear regression. Authors discuss the advantage of those algorithms the in big data system. (Neyshabouri et al. 2016) present an algorithm for nonlinear regression in big data system based on lexicographical splitting graph (Wang et al.,2015) this algorithm divide n data into 2n possible partitions to construct sequence piecewise linear model, and combines them (Willems et al.1996) proposed Cover's theorem(Cover ;1965), which can transform training data set non linearly separable in tanning set linearly separable. This work divided data set into tanning data set and test data set the proposed algorithm to generate a huge number of (104 -106) of random feature intermediate is given predictor matrix for the training data set, and they use training test data sets to choose predictive intermediate features by regularized linear or logistic regression.

K-means algorithm

The k-means algorithm takes into account k input parameter, and partition a set of attributes in K clusters. Cluster similarity is measured about the average value of objects in a cluster, which can be considered as the cluster's centroid or center of gravity (Han ;2011)

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

Algorithm 1: K-means algorithm

Input: k,the number of cluster. D,data containing n attribute.
Output: A, a set of k cluster
1 Choose k attribute from D
2 repeat
3 (re)assign each attribute to the cluster to which the attribute is the most similar, based on the mean value of the attributes in the cluster
4 update the cluster means, i.e., calculate the mean value of the attributes for each cluster;
5 until No change

k-means algorithm calculate the square error criterion:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

Where E is the sum of the square error for all attributes,p is the point in space representing a given

attribute, and m_i is the mean of cluster C_i.

PROPOSITION

Linear model for curve regression

Let X = {x_1, x_2, ... x_n} data set of curve model divided into m sub data set {x^1, x^2, ...x^m} in big data architecture. The first step in our mathematic model is convert the curve model into linear model by linearization as we presented in Table 2, for each sub data set {x^1, x^2,...x^m} convert in linear model {Z^1, Z^2, ...Z^m},where Zi={z^i_0; z^i_1; z^i_2...z^i_l} The general model of sub data i expressed by y^i a^i_j ,z^i and b^i.

$$y^i = a^i_0 z^i_0 + a^i_1 z^i_1 + ...a^i_l z^i_l + b^i.$$

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

(7)

This step can returns the vectors{v_1,v_2, ...v_m} Where vi = (a^i_0; a^i_1; a^i_2; :::a^i_l; bi)

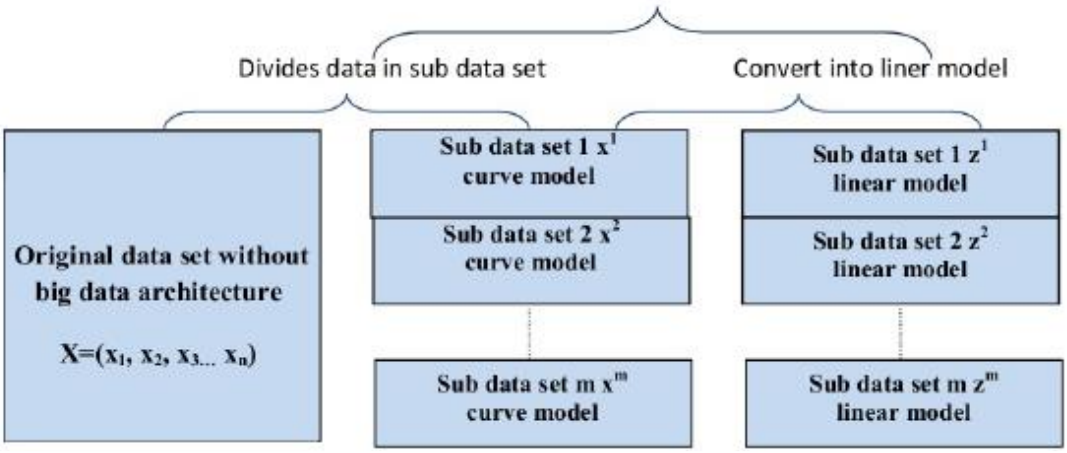


Fig. 3 – Dividing Data and convert into linear model.

Map algorithm

Curve model divided into m nodes in big data architecture. Map algorithm can transform each data node, into a linear model, as we describe in 3.1.

Algorithm 2: Map Algorithm

Input: Sub data set fore node i and x^i, where $x^i = \{x^i_1, x^i_2, ...x^i_l\}$
Output: (i,v^i) ,where i represent a node and v^i linear models parameters,v^i=$(a^i_1, a^i_2, ...a^i_m, b^i)$

1 $y = f(x^i_1, x^i_2, ...x^i_l)$ ▷ f function can transform curve model into linear
2 Output(i,v^i) ▷ Returns for each node i the vector v^i

Select clusters by Reduce k-means algorithm

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

After determined the linear regression of each sub data set in node i, we apply Reduce k-means algorithm, to performs hard clustering, each linear model assigned only to one cluster, that can select bests linear models. The Reduce k-means algorithm process as follows. First, it randomly generates k from v^i = {1 : : : i = m}, each of which initially represents a cluster mean or center.For each of the remaining v^i {i = 1... i = m}, a v^i is assigned to the cluster C_j to which it is the most similar, based on the distance between vi and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges.

Algorithm 3: Reduce K-meansAlgorithm

Input: A set of (i, v^i) ={$(1, v^1), (2, v^2), \ldots, (m, v^m)$}, Where i represent a node and $v^i = \{a_1^i, a_2^i, \ldots, a_l^i, b^i\}$ parameters of linear models.

Output: {$(m, (C_1, C_2, \ldots, C_k))$}, Where m the number of node and (C1,C2...Ck)list of cluster.

initialize k centroids:$\{\mu_1^t, \mu_2^t, \ldots, \mu_k^t\}$
$t \leftarrow 0$
$i \leftarrow 1$
repeat
  $t \leftarrow t + 1$
  $C_j \leftarrow \emptyset$
  while $i <= m$ do
    $J^* \leftarrow argmin_i \|v^i - \mu_k^t\|^2$ ▷ Assign v^i to closest centroid
  $C_{j^*} \leftarrow C_{j^*} \cup v^i$
  $i \leftarrow i + 1$
  while $l <= k$ do
    $\mu_l^t \leftarrow \frac{1}{C_l} \sum_{i=1}^{m} (v^i)$
until $\sum_{l=1}^{k} (\|\mu_l^t - \mu_l^{t-1}\|)$
Output(m,$(C_1, \ldots, C_k)$) ▷ Returns clusters$(C_1, \ldots, C_k)$for m nodes

Work flow of our architecture

The work flow architecture Figure 4, presents data nodes (Data Node1, Data Node2..., DataNodem), and algorithms executes on it.The Map algorithm(Map algo1,Map algo2,...Map algom) execute in each node in order to extract linear model.In the reduce phase algorithm (Reduce algo) extracts K clusters (C_1,C_2...C_k).

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

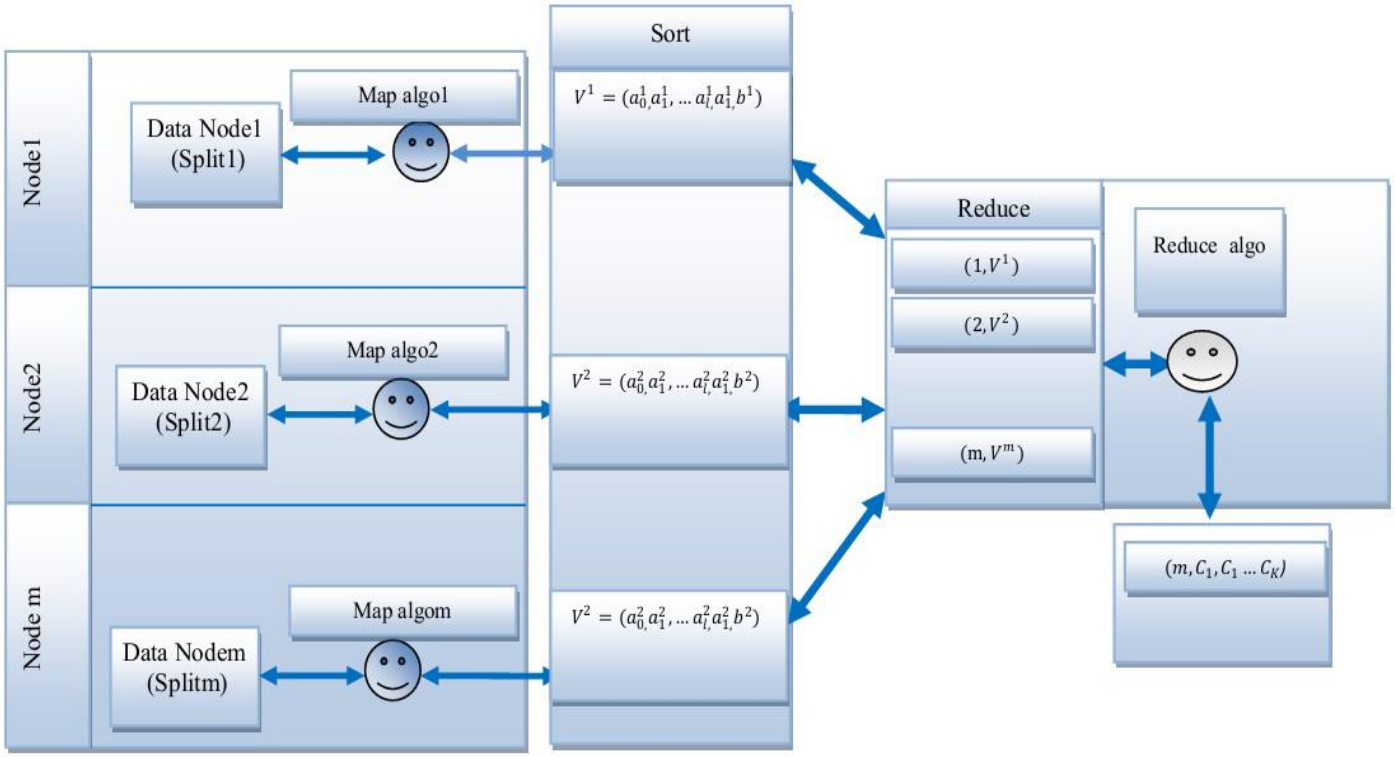


Fig. 4 – Work flow architecture.

Table 3 - Results of linear models.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

1 Node(m=1)		2 Nodes (m=2)				3 Nodes (m=3)			
b_1^1	a_1^1	b_1^2	a_1^2	b_2^2	a_2^2	b_1^3	a_1^3	b_2^3	a_2^3
-6.127315	0.037125	-6.098598	0.037298	-6.151394	0.036893	-5.977548	0.036779	-6.429044	0.039286

3 Nodes (m=3)		4 Nodes (m=4)							
b_3^3	a_3^3	b_1^4	a_1^4	b_2^4	a_2^4	b_3^4	a_3^4	b_4^4	a_4^4
-5.995022	0.035323	-6.083422	0.0.037931	–6.151577	0.036973	-6.190879	0.037659	-6.11194	0.036072

5 Nodes (m=5)									
b_1^5	a_1^5	b_2^5	a_2^5	b_3^5	a_3^5	b_4^5	a_4^5	b_5^5	a_5^5
-6.594134	0.041134	-5.527216	0.033683	-6.387652	0.038618	-6.481455	0.040052	-5.846709	0.034605

6 Nodes (m=6)									
b_1^6	a_1^6	b_2^6	a_2^6	b_3^6	a_3^6	b_4^6	a_4^6	b_5^6	a_5^6
-6.464922	0.04014	-5.524988	0.032566	-6.527205	0.040498	-6.115041	0.037279	-6.187981	0.036382

6 Nodes(m=6)		7 Nodes (m=7)							
b_6^3	a_6^3	b_1^7	a_1^7	b_2^7	a_2^7	b_3^7	a_3^7	b_4^7	a_4^7
-6.158437	0.036117	-6.556825	0.040673	-5.889929	0.037013	-5.72548	0.032986	-6.702826	0.042033

7 Nodes(m=7)						8 Node (m=8)			
b_5^7	a_5^7	b_6^7	a_6^7	b_7^7	a_7^7	b_1^8	a_1^8	b_2^8	a_2^8
-6.048399	0.036412	-6.042611	0.036109	-6.116011	0.035855	-6.649196	0.041541	-5.662329	0.035355

8 Nodes(m=8)									
b_3^8	a_3^8	b_4^8	a_4^8	b_5^8	a_5^8	b_6^8	a_6^8	b_7^8	a_7^8
-5.719456	0.036412	-6.704018	0.041565	-6.463367	0.039195	-5.968078	0.036449	-6.229845	0.037863

8 Node(m=8)		9 Nodes (m=9)							
b_8^8	a_8^8	b_1^9	a_1^9	b_2^9	a_2^9	b_3^9	a_3^9	b_4^9	a_4^9
-5.979982	0.03397	-6.437492	0.040294	-6.251101	0.039303	-5.427199	0.032275	-6.24298	0.036775

9 Nodes(m=9)									
b_5^9	a_5^9	b_6^9	a_6^9	b_7^9	a_7^9	b_8^9	a_8^9	b_9^9	a_9^9
-6.528791	0.040701	-6.575706	0.040892	-5.865148	0.034322	-6.348205	0.039533	-5.746428	0.031455

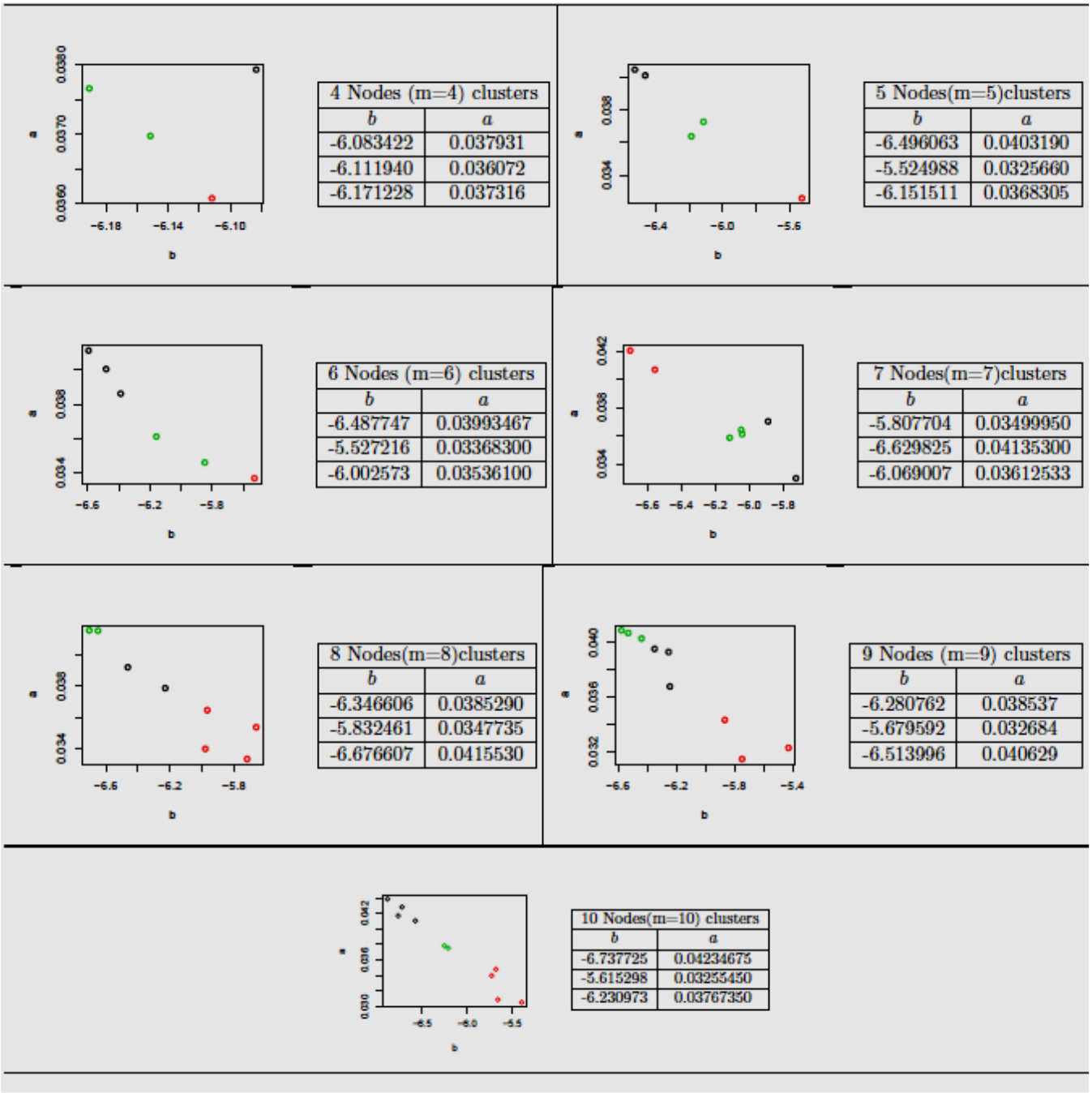
10 Nodes (m=10)									
b_1^{10}	a_1^{10}	b_2^{10}	a_2^{10}	b_3^{10}	a_3^{10}	b_4^{10}	a_4^{10}	b_5^{10}	a_5^{10}
-6.724393	0.042816	-6.252385	0.037817	-5.865148	0.034781	-5.393673	0.030529	-6.885722	0.043841

Node10									
b_6^{10}	a_6^{10}	b_7^{10}	a_7^{10}	b_8^{10}	a_8^{10}	b_9^{10}	a_9^{10}	b_{10}^{10}	a_{10}^{10}
-6.20956	0.03753	-6.574634	0.041044	-5.728508	0.03399	-6.766149	0.041686	-5.657988	0.043841

Apply k-meas algorithm

The second step of our proposition, apply the Reduce k-means algorithm. We select 3 clusters (k=3). Our algorithm takes linear models parameters extracted from Map Algorithm 2 and, construct 03 clusters. .For example in node5,C1 = (-6.496063, 0:0403190),C2 = (-5.524988, 0:0325660) and C3 = (-6.151511, 0.0368305).The result of (m=4...m=10) in figure6.

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48



4 Nodes (m=4) clusters	
b	*a*
-6.083422	0.037931
-6.111940	0.036072
-6.171228	0.037316

5 Nodes(m=5)clusters	
b	*a*
-6.496063	0.0403190
-5.524988	0.0325660
-6.151511	0.0368305

6 Nodes (m=6) clusters	
b	*a*
-6.487747	0.03993467
-5.527216	0.03368300
-6.002573	0.03536100

7 Nodes(m=7)clusters	
b	*a*
-5.807704	0.03499950
-6.629825	0.04135300
-6.069007	0.03612533

8 Nodes(m=8)clusters	
b	*a*
-6.346606	0.0385290
-5.832461	0.0347735
-6.676607	0.0415530

9 Nodes (m=9) clusters	
b	*a*
-6.280762	0.038537
-5.679592	0.032684
-6.513996	0.040629

10 Nodes(m=10) clusters	
b	*a*
-6.737725	0.04234675
-5.615298	0.03255450
-6.230973	0.03767350

Fig. 6 - Results of k-means clustering for nodes (k=3).

Revista Cubana de Ciencias Informáticas
Vol. 14, No. 2, Abril-Junio, 2020
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 34-48

DISSCUTION OF OUR APPROACH

Our approach is a complete approach toward regression problem in big data; it covered the mathematic models such as (Jun et al.,2015; Ma et al., 2015; Neyshabouri et al 2016) works, and MapReduce algorithm and architecture like (Oancea et al.,2015). Moreover, our approach combines between to important problem of data mining, regression, and machine learning problems. Map algorithm can solve the regression problem of curve regression; it can convert curve model into linear model and Reduce k-means algorithm can represent the clustering problem. Big data architecture composes by various nodes; each node returns linear model. Consequently, reduce k-means algorithm select the best k-clusters wich can describe linear models.

CONCLUSION AND FUTURE WORK

In this paper, we have proposed curve regression in big data system.Data in our architecture is divided into sub data, each sub data assigned to node, the first algorithm in our approach converts the curve model into linear model, each node convert its sub data into linear model. In the second step, we apply k-means algorithm for each node in order to extract clusters. We validate our approach by UniversalBank data set; we calculate linear models parameters and obtain 03 clusters for each node. Our approach combine the regression with clustering problem in big data architecture, the result extracted from Map algorithm input into Reduce k-means algorithm to select the clusters which can better represent the regression model.

REFERENCE

Bollobás, Béla. Linear analysis. Cambridge: Cambridge University Press, 1990.10p.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers,1965: (3), 326-334.

Dean, J., And Ghemawat, S. MapReduce: a flexible data processing tool. Communi- cations of the ACM, 2010. 53(1): p.72-77.

Golberg, Michael A., And Hokwon A. Cho. Introduction to regression analysis. WIT press, 2004.3p.

Han, J., Pei, J., And Kamber, M. Data mining: concepts and techniques. Elsevier,2011.

Jun, S., Lee, S. J., And Ryu, J. B. A Divided Regression Analysis for Big Data. Statistics, 2015:9(5).

Krishna, K, Open source implementation of MapReduce,2019. [On line]. Teck Kaizen 2010, [Accessed on: April,2019] Disponible en: http://kktechkaizen.blogspot.com/2012/07/apache-hadoop-open-source-mapreduce.html.

Ma, P., And Sun, X, Leveraging for big data regression. Wiley Interdisciplinary Re- views: Computational Statistics, 2015:7(1), p.70-76.

Naoui, M. A., Mcheick, H., Kazar, O. Mobile Agent approach based on mo- bile strategic environmental Scanning using Android and JADELEAP systems. In Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on IEEE,2014: p.1-7.

Neyshabouri, M. M., Demir, O., Delibalta, I., And Kozat, S. S. Highly efficient non- linear regression for big data with lexicographical splitting. Signal, Image and Video Processing, 2016, p.1-8.

Oancea., B..Linear Regression With R And HADOOP.International Conference : CKS Challenges of the Knowledge Soc;2015, p1007.

Shafer, J., Rixner, S., And Cox, A. L. The hadoop distributed filesystem: Balancing portability and performance. In Performance Analysis of Systems and Software (ISPASS), 2010: p. 122-133.

V.Martha, W. Zhao, Xiaowei Xu,. h-MapReduce: A Framework for Workload Balancing in MapReduce. IEEE 27th International Conference on Advanced Information Networking and Applications 2013: p.637-644.

Wang, Y., Li, Y., Xiong, M., And Jin, L. Random Bits Regression: a Strong General Predictor for Big Data. arXiv preprint arXiv,2015.

Willems, F. M., Shtarkov, Y. M., And Tjalkens, T. J. Context weighting for general finite-context sources. IEEE transactions on information theory, 1996:42(5),p. 1514-1520.

Conflicto de interés

No existe conflicto de interés con este trabajo

Contribuciones de los autores

El primer autor contribuyo el 60% del trabajo y el segundo y tercera el 40%.