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 Naoui¹: https://orcid.org/0000-0003-1653-531X
Brahim Lejdel²: https://orcid.org/0000-0003-1779-0689
Mouloud Ayad³*: https://orcid.org/0000-0001-9858-8612

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

---

**Introducción**

Regression analysis (Golberg et al., 2004) is a statistical methodology describe the relationship between variables attributes. 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:

\[ y = ax + b \]  \hspace{1cm} (1)

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) \}

\[ b = y^- - x^- \]  \hspace{1cm} (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) \}.

\[ a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \]  \hspace{1cm} (3)
Multiple linear regression

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

\[ y = a_x + a_2x_2 + ... + a_lx_l + b \] (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>
<thead>
<tr>
<th>Model</th>
<th>General expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power Model</td>
<td>( y = bx^a )</td>
</tr>
<tr>
<td>Exponential Model</td>
<td>( y = be^ax )</td>
</tr>
<tr>
<td>Logistic model</td>
<td>( y = l - e^{(ax+b)} )</td>
</tr>
<tr>
<td>Polynomial model</td>
<td>( y = a_nx^n + a_{n-1}x^{n-1} + ... + a_1x + b )</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>General expression</th>
<th>Linearization</th>
<th>( y' ) =</th>
<th>( x' = )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y = bx^a )</td>
<td>( \log(y) = a\log(x) + \log(b) )</td>
<td>( y' = \log(y) )</td>
<td>( x' = \log(x) )</td>
</tr>
<tr>
<td>( y = be^{ax} )</td>
<td>( \ln(y) = ax + \ln(y) )</td>
<td>( y' = \ln(y) )</td>
<td>( x' = x )</td>
</tr>
<tr>
<td>( y = l - e^{(ax+b)} )</td>
<td>( \ln(y) = ax + b - \ln(l) )</td>
<td>( y' = \ln(y) )</td>
<td>( x' = x )</td>
</tr>
<tr>
<td>( y = a_nx^n + a_{n-1}x^{n-1} + ... + a_1x + b )</td>
<td>( y' = a_1x_1 + a_2x_2 + ... + a_nx_n + b )</td>
<td>( x'_n = x^n, x'_2 = x^2, x'_1 = x^1 )</td>
<td></td>
</tr>
</tbody>
</table>
Big data MapReduce Algorithms

MapReduce \cite{Dean2010} 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 \cite{Shafer2010,Martha2013}. Hadoop \cite{Krishna2010} 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. Validation and results of our proposition of UniversalBank 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 \cite{Jun2015,Oancea2015,Ma2015,Neyshabouri2016}. Several works oriented to propose mathematic approaches for regression in big data such as data \cite{Jun2015, Ma2015, Neyshabouri2016}. Other geared to proposes MapReduce algorithms and its implementations in big data system like \cite{Oancea2015}

Linear model

\cite{Jun2015} presented a divided regression analysis using multiple linear where regression form is:

\[ y = a_1x_1 + a_2x_2 + ...a_mx_m + b \]
Authors use random sampling data to divide big data into sub samples, they consider all attributes have an equal chance to be selected in the sample Fig 1. (Oancea et al. 2015) 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.

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 $2^n$ 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 ($10^4$ -$10^6$) 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)

```
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 \( X = \{x_1, x_2, \ldots, x_n\} \) data set of curve model divided into \( m \) sub data set \( \{x^1, x^2, \ldots, x^m\} \) 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 \( \{x^1, x^2, \ldots, x^m\} \) convert in linear model \( \{Z^1, Z^2, \ldots, Z^m\} \), where \( Z_i = \{z^1_i; z^2_i; \ldots; z^m_i\} \) The general model of sub data \( i \) expressed by

\[
y^i = a^i_0 z^i_0 + a^i_1 z^i_1 + \ldots + a^i_l z^i_l + b^i.
\]

This step can returns the vectors \( \{v_1, v_2, \ldots, v_m\} \) Where \( v_i = (a^i_0; a^i_1; a^i_2; \ldots; a^i_l; b^i) \)

**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.
Select clusters by Reduce k-means algorithm

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 \( v^i \) and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges.
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 (C₁,C₂,...Cₖ).

Fig. 4 – Work flow architecture.
Table 3 - Results of linear models.

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Node 1 (m=1)</th>
<th>Node 2 (m=2)</th>
<th>Node 3 (m=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>b_1^1 = -6.127315, a_1^1 = 0.037126</td>
<td>b_1^2 = -6.098598, a_1^2 = 0.037298</td>
<td>b_1^3 = -5.977548, a_1^3 = 0.036893</td>
</tr>
<tr>
<td>2</td>
<td>b_2^1 = -5.995022, a_2^1 = 0.035323</td>
<td>b_2^2 = -6.083422, a_2^2 = 0.037931</td>
<td>b_2^3 = -5.769879, a_2^3 = 0.037659</td>
</tr>
<tr>
<td>3</td>
<td>b_3^1 = -6.691436, a_3^1 = 0.041134</td>
<td>b_3^2 = -6.563271, a_3^2 = 0.035838</td>
<td>b_3^3 = -5.561455, a_3^3 = 0.040862</td>
</tr>
<tr>
<td>4</td>
<td>b_4^1 = -6.464922, a_4^1 = 0.040778</td>
<td>b_4^2 = -6.524988, a_4^2 = 0.035506</td>
<td>b_4^3 = -6.115041, a_4^3 = 0.037279</td>
</tr>
<tr>
<td>5</td>
<td>b_5^1 = -6.158437, a_5^1 = 0.036117</td>
<td>b_5^2 = -6.566825, a_5^2 = 0.040673</td>
<td>b_5^3 = -5.725484, a_5^3 = 0.032968</td>
</tr>
<tr>
<td>6</td>
<td>b_6^1 = -6.048399, a_6^1 = 0.036412</td>
<td>b_6^2 = -6.042611, a_6^2 = 0.036109</td>
<td>b_6^3 = -6.049196, a_6^3 = 0.041341</td>
</tr>
<tr>
<td>7</td>
<td>b_7^1 = -5.710456, a_7^1 = 0.036412</td>
<td>b_7^2 = -6.740118, a_7^2 = 0.041565</td>
<td>b_7^3 = -5.968078, a_7^3 = 0.036449</td>
</tr>
<tr>
<td>8</td>
<td>b_8^1 = -5.979082, a_8^1 = 0.033977</td>
<td>b_8^2 = -6.437402, a_8^2 = 0.040294</td>
<td>b_8^3 = -5.427109, a_8^3 = 0.032275</td>
</tr>
<tr>
<td>9</td>
<td>b_9^1 = -6.598791, a_9^1 = 0.040701</td>
<td>b_9^2 = -6.575076, a_9^2 = 0.040892</td>
<td>b_9^3 = -6.348205, a_9^3 = 0.039633</td>
</tr>
<tr>
<td>10</td>
<td>b_10^1 = -6.724393, a_10^1 = 0.042816</td>
<td>b_10^2 = -6.253285, a_10^2 = 0.037817</td>
<td>b_10^3 = -5.393673, a_10^3 = 0.030602</td>
</tr>
</tbody>
</table>

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 node 5, C1 = (-6.496063, 0.0403190), C2 = (-5.524988, 0.0325660) and C3 = (-6.151511, 0.0368305). The result of (m=4...m=10) in figure 6.
Fig. 6 - Results of k-means clustering for nodes (k=3).
Discussion 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


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


V.MARTHA, W. ZHAO, XIAOWEI XU,. h-MapReduce: A Framework for Workload Balancing in

Conflicto de interés

No existe conflicto de interés con este trabajo

Contribuciones de los autores

Mohammed Anouar Naoui: Contribuyó en el enfoque propuesto que abarca arquitectura y algoritmo.
Brahim Lejdel: Contribuyó en la supervision y mejora de la arquitectura.
Mouloud Ayad: Contribuyó en la co-supervisión y mejora del algoritmo.

Financiación

No se obtuvo financiamiento para realizar este trabajo