Enhancement of Big Image Processing Using Naive Bayes-based Logistic Regression

R.N.V.Jagan Mohan
Professor,
Department of Computer Science and Engineering,
Swarmandhra College of Engg & Tech, Narsapuram-534 275.
E-mail:mohanrnvj@gmail.com.

Abstract: Nowadays, the digital substance is generated with exponential pace due to the abundant use of scanners and video cameras. As the number of digital images is ever increasing, the aim of storing and processing the colossal database has become a hard task. According to the recent surveys, most of the digital data available today is unstructured such as video, audio, images or text. Classification is playing a very essential role for offline/online based image processing. Due to Large-scale geological image data, classification at the level of pixels is necessitated. In this paper, we proposed a classification approach on raster based spatial image data using Naive Bayes-based Logistic Regression analysis. In this study, the images of natural resources in raster data form are used to get thematic maps. The main aim of this research is to check whether all the objects are conditionally independent of the given class, which can be used to get the more bias and lower variance. Finally, the experiments are conducted on spatial big images in order to increase the efficiency of the process for this system.

Keywords: Big Image, Conditional Independent, Logistic Regression, Naive Bayes, Spatial Image.

I.INTRODUCTION

Extracting the frames from videos has features to improve in tackling the wide-ranging experiments of huge data, i.e., Bigdata. Bigdata is a term of datasets that describes large volumes of data, whose measure, variety, and complexity needed for new architecture, techniques, algorithms, and analytic to manage it and extract value and hidden knowledge from it. The amount of images has been uploaded to the internet is rapidly increasing, such as, Facebook users uploading over 2.5 billion new photos in every month (Facebook, 2010) [2]. However, make use of this data are severely lacking for transforming into several applications. At present, computer vision applications uses a small number of input images because of the difficulty in acquiring computational resources and storage options for large amounts of data projected in (Guo, 2005)[3] and in (White et al., 2010)[8]. As such, the development of vision applications suggested by Ghemawat and Gobioff, 2003[4], have limited that use a large set of images. Many image processing and computer vision procedures are applicable to large-scale data tasks. It is often desirable to run these algorithms on large data sets, for example, larger than one-terabytes that are currently limited by the computational power of one computer (Guo, 2005)[3]. Typically, these tasks performed on a distributed system by dividing the task across one or more of the following feature algorithm parameters, such as, images or pixels (White et al., 2010a) [8] (White et al., 2010b)[9]. Performing tasks across a particular parameter is incredibly similar and can often be parallel. Face detection and landmark classification are examples of such algorithms, well defined in (Li and Crandall, 2009) [5] and in (Liu et al. 2009) [6]. The ability to parallelize such tasks allows for scalable, efficient execution of resource intensive applications.
By using distributed processing, we can do computing, information access, information exchange across multiple platforms connected by computer networks. With the increase of high-speed communication networks and powerful computers, network-based computing systems have proven to be an economical and efficient alternative to supercomputers. In distributed computing systems, a critical concern is to efficiently partition and schedules the tasks between available processors in such a way that the overall processing time of the submitted tasks is at minimum. Image data is the right candidate to exploit data parallelism because they are inherently divisible in nature. Therefore, distributed image processing is used for large image data processing.

The rest of the paper is organized as follows: Unstructured Frames using Naïve-Bayes approach is discussed in Section 1. The details of proposed work and discussion of various parameters, which affect its performance, are discussed in Section 2. Section 3 deals with Logistic Regression. The experimental results are highlighted in Section 4. Section 5 deals with the conclusion.

II. UNSTRUCTURED FRAMES

Unstructured data like photos and other multimedia data cannot be categorized or analyzed numerically. The most important stages of the big data sequence are data analysis, the goal of which is to extract useful values, suggest conclusions and/or support decision-making. First, we discuss the purpose and classification metric of data analytics. Second, we review the application evolution for various data sources. Here, we discuss about the classification and obtain the picture file, i.e., un-structure file and it is not comparing with other files. This file can be extracts into various input frames. These frames are used as input to unseen structure in unlabeled frames that is categorized into homogeneous groups for the purpose of identification. These big frame profiles are very important for invariant systems. Naïve-Bayes classification technique is used for an uncomplicated and text classification. This algorithm can be viewed as a linear classifier, that means each classifier is corresponds to a hyper-plane. The classification approach has applied for the given input image (record) I to classify, the general approach is to output the class Ci whose probability of occurrence P(Ci|I) is maximized. To estimate the value of P(Ci|I), this classifier naively assumes that the attributes of I are independent of each other called as Naïve-Bayes(Ng et al., 2002) [7]. The independence of the given record of values P(Ci|I) to be computed is as follows:

\[ P(C_i|I) = \frac{P(I \cap C_i)}{P(I)} = \frac{P(C_i)P(I)}{P(C_i)P(I)} = P(I|C_i)P(C_i) \]  

III. BASIC ALGORITHM

The notion of the system image data is to produce the GIS based picture of the Big Image algorithm using Remote Sensing. In this, a vector representation of raster images in ‘of space’ gives to the exact results, low computational costs and better response time. To study of this work is based on creating vector model from vectorization (raster-to-vector conversion) and storing into repository. For this purpose of image analysis, the information, which is a stored in repository i.e., very similar nearest neighbor pixels individually. It is an inefficient that’s why we use optimization to increase in the number of records intern increases cost of analysis. To optimize the storage and cost analysis a new method is proposed.
The following are the phases of the proposed algorithm:
1) Initially, load the LISS-IV (Satellite Image) GIS images into GIS Database.
2) Convert GIS Image data into Vector Data.
3) Form the signatures and generate vector data.
4) Use the pixel classifications and label the signature data.
5) Finally, merge the pixels using similarity-based distance measures approach used for creating thematic maps.

![Image 1](image1.png)

Figure 2. Big Image Data Processing

Initially, Raster data (LISS-IV satellite image in TIFF format) and reference data are loaded into the spatial database. Then, the raster data are pre-processed and converted into vector data. Later signatures are generated based on evolved vector data and the reference vector data. After that, the classes are defined based on the signature pixels, from the signature data table. For the defined classes, the classification is performed and tabulated in a classification table. In this system, pixels are classified and merging of pixels of the same class is done. This can be referred as post-processing and thematic maps are generated based on this approach.

IV. LOGISTIC REGRESSION

This method is used for learning functions of the form $f$: $I \rightarrow C$, $P(I|C)$, where ‘I’ is discrete-value, and $I=\{I_1, I_2... I_n\}$ is any vector containing discrete or continuous variables, and where ‘I’ is a Boolean variable, in order to simplify the notation. In this regard, we choose our behavior in the case where ‘I’ take on any finite number of discrete values (Ng, A.Y and Jordan, M.I, 2002) [7].

This model undertakes a parametric form for the distribution $P(I|C)$, then directly estimates its parameters from the input data suggested by (Ng, A.Y. and Jordan, M.I, 2002) [7]. The parametric model assumed by Logistic Regression in the case where I is Boolean is

$$P(I = 1|C) = 1 < \exp \left( W_0 + \sum_{i=1}^{n} W_i I_i \right)$$

Then, taking the natural log in both sides, we have a linear classification rule that assigns labels $I=0$ if C satisfies, otherwise assign $I=1$.

$$0 < \exp \left( W_0 + \sum_{i=1}^{n} W_i I_i \right)$$
This method is used on training data directly estimating $P(I|C)$, in deference to Naïve Bayes. Logistic regression is often referred to as a discriminative classifier because we can view the distribution $P(I|C)$ as directly disseminating the value of the target value ‘I’ for any given instance C. The Naïve Bayes is a learning algorithm with greater bias, but lower variance than logistic regression. If bias is an appropriate for the given actual data, then Naïve Bayes will be preferred or else Logistic regression will be preferred.

V.EXPERIMENTAL RESULTS
In the experimental results, we calculate for various pixels of satellite images in object direction using regression. In Table 4.1, the recognition values are obtained and are tabulated. As shown in Figure 4.2, it exposes the graphical observation of the highest recognition level of the input image.

![Figure 1. Satellite Image](image1.png)

![Figure 2. Process of Satellite Image](image2.png)
In the above Table I, it is observed that if we take first option, pixel by pixel, then the rows are affected heavily and execution time is less, where as in the second option merging pixels, the rows are affected low and execution time is high.

![Figure 3. Recognize Location of Satellite Image 1](image1)

![Figure 4. Recognize Location of Satellite Image 2](image2)

<table>
<thead>
<tr>
<th>Method</th>
<th>Rows Effected</th>
<th>Execution Time (Ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PixelByPixel(PP)</td>
<td>1307095</td>
<td>9063</td>
</tr>
<tr>
<td>Merging Pixels(MP)</td>
<td>539682</td>
<td>191147</td>
</tr>
</tbody>
</table>

In the above Table I, it is observed that if we take first option, pixel by pixel, then the rows are affected heavily and execution time is less, where as in the second option merging pixels, the rows are affected low and execution time is high.

![Figure 5. Pixels loaded in the Database and Execution Time in Loading Pixels](image3)

<table>
<thead>
<tr>
<th>Method</th>
<th>Execution Time (Ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PixelByPixel</td>
<td>23525</td>
</tr>
<tr>
<td>Merging Pixels</td>
<td>7286</td>
</tr>
</tbody>
</table>
In the above Table II, it is observed that there are two options; one is pixel by pixel, if we take this option then time of indexing-based execution is high. Whereas in the second option, merging of pixels; if we take this option then execution time is low.

![Figure 6. Times for Indexing Pixels in Database](image)

Table III

<table>
<thead>
<tr>
<th>Class Name</th>
<th>Area (Hectors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water Resources</td>
<td>214.17</td>
</tr>
<tr>
<td>Aquaculture</td>
<td>1869.64</td>
</tr>
<tr>
<td>Agriculture</td>
<td>579.69</td>
</tr>
<tr>
<td>Built-up Land</td>
<td>191.2</td>
</tr>
</tbody>
</table>

![Figure 7. Total Area in Hectors](image)

VI. CONCLUSION

Statistical estimators of functions or of conditional distributions are used with the help of Naïve Bayes and Logistic regression. They estimate $P(I|C)$ from a sample of training data by bias and expected variance because Naïve Bayes is a learning algorithm with greater bias, but lower variance than logistic regression. The classification is on Raster based GIS image data and this image is taking from Upputeru Reservoir to find a thematic map. Initially we took its raster data. The experiments are conducted on big images is to increase the efficiency of the process for this system.
REFERENCES