Naïve Bayes & Logistic Regression on Big Data: A Performance Analysis

Sharon Susan Jacob¹, Dr. R. Vijayakumar ²

¹Research Scholar, School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India.
²Professor, School of Computer Sciences, Mahatma Gandhi University, Kottayam, Kerala, India.

Abstract

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. Big data analytics is the practice of analysing huge data sets to discover unknown patterns, hidden correlations, preferences and other useful information. Twitter is one of the famous social networks worldwide and main source of Bigdata. Twitter tweets and other social media posts are some of the central source of unstructured data. To find meaningful information from unstructured data is very difficult. With the help of classification technique it is possible to change unstructured data into organised form. In this work, twitter data was taken as the dataset. For training data, supervised machine learning algorithms like Naïve Bayes and Logistic Regression are used. For analysing the data, Python programming was used. The result shows that Naïve Bayes classifier yielded more classification accuracy than Logistic Regression classifier.

Keywords: BigData, NaiveBayes, LogisticRegression, Python.

INTRODUCTION

Big data is being generated by everything around us and is not a single technology but a combination of old and new technologies that help companies gain actionable insight. So the analysis in big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. Big data is typically broken down by three characteristics: Volume, Velocity and Variety. Volume Refers to the amount of Data that is getting generated. Velocity point out the speed at which data is getting generated. Variety indicates the different types of data that is getting generated.

Naïve Bayes is a probabilistic classifier based on applying Bayes’ theorem with strong independence assumption that the presence of one feature in a class does not depend on the presence or absence of another feature. Logistic Regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, fraud etc.) or 0 (FALSE, failure, not-fraud, etc.). The goal of logistic regression is to find the best fitting model to describe the relationship between dependant variable and a set of independent variables.

To find meaningful information from unstructured data is very difficult. But with the help of classification technique it is possible to change unstructured data into organised form. In this paper, the main focus is on comparative performance of Naïve Bayes and Logistic Regression in the analysis of Big data.

RELATED WORKS

Ahmed Fuad Mohammed et al. [1] in their work A Review of Big Data Environment and Its Related Technologies gives a detailed picture of big data and its background. The paper presents different categories of Big Data, its own categories and complexities. The paper also reviews existing technologies on Big Data. Moreover the paper consist of comparison between Big Data and Cloud Computing Platforms.

Nakov et al. [2] presented an evaluation of a Semantic analysis task in which two tasks were performed – the sentiment analysis of the social media content and the natural language processing of the text. It compares various approaches for analysing the social media content such as twitter.

Qiaowei Jiang et al. [3] proposed a high efficient method called deep feature weighting Naive Bayes. The deep feature weighting Naive Bayes applied to Chinese text classifiers obtained a better performance than ordinary feature weight Naive Bayes.

Geetika Gautam and Divakar Yadav [4] implemented supervised algorithms such as SVM, Naïve Bayes and maximum entropy to classify the twitter dataset based on sentiments. The results obtained from algorithms are compared based on their relative performances on three parameters namely: accuracy, precision and recall. A comparative measurement is taken on the classification using supervised learning algorithms and the semantic analysis.

METHODOLOGY

The data set after pre-processing and feature extraction, divided into Training data and Test data. The training data is given to machine learning and after that the test data is given. The algorithms used are Naïve Bayes and Logistic Regression and the language used is Python. The performance of these algorithms are analysed.
Data set

There are large amount of data, increasing every day because of social media. Twitter is one of the social media site which gives an opportunity to people to express their ideas and opinions about a particular topic. Twitter tweets and other social media posts are some of the central source of unstructured data. So twitter data is taken as the Data set.

Pre-Processing

The pre-processing of the data is a very important step as it decides the efficiency of the other steps down in line and it involves removal of unimportant features from the data. The steps involved should aim for making the data more machine readable in order to reduce ambiguity in feature extraction. Below are a few steps used for pre-processing of tweets:

- **Removal of re-tweets**
- **Converting upper case to lower case**: In case we are using case sensitive analysis, we might take two occurrence of same words as different due to their sentence case
- **Stemming and Lemmatization**: Stemming and Lemmatization are two essential morphological processes of pre-processing module during feature extraction.

Feature Extraction

The quality and quantity of features is very important as they are important for the results generated by the selected model. Selection of useful words from tweets is feature extraction.

- **Bag of Words**: This technique involves the following tasks: 1. Tokenizing strings and giving an integer id for each possible token, for instance by using white-spaces and punctuation as token separators. 2. Counting the occurrences of tokens in each document. 3. Normalizing and weighting with diminishing importance of tokens that occur in the majority of samples / documents.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: The tf-idf transform is used in order to re-weight the count to shadow the frequencies of rarer yet more interesting features into floating point values suitable for usage by a classifier. Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency.

Algorithms applied for training the data set

For preparing and training the data, the following the 2 algorithms are implemented.

- **Naïve Bayes**: Naïve Bayes is a simple model which works well on text categorization. For tweets a multinomial Naïve Bayes model can be used

\[
(c^*) = \arg \max_c PNB(c|d) \\
P NB(c|d) := (c) (f|c) n_c(b)
\]

Class \( c^* \) is assigned to tweet \( d \). \( f \) represents a feature and \( n_c(b) \) represents the count of feature \( f \), found in tweet \( d \). There are a total of \( m \) features.
Parameters $P(c)$ and $P(f|c)$ are obtained through maximum likelihood estimates. $P(c)$ is prior probability.

- **Logistic Regression (LR):** Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

$$\logit(p) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

where $p$ is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds.

$$\text{odds} = \frac{p}{1-p}$$

$$\logit(p) = \ln \left( \frac{p}{1-p} \right)$$

**ANALYSIS OF THE TESTING DATA**

Twitter sentiment analysis was done using the language Python. Python is a high level, interpreted programming language and is very popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks and provides a large standard library, which can be used for various applications. In using the Language Python, the following packages are used:

**SCIKIT-LEARN**

The Scikit-learn project started as scikits.learn, a Google Summer Code project. It is a powerful library that provides many machine learning classification algorithms, efficient tools for data mining and data analysis. Below are various functions that can be performed using this library:

- Classification: Identifying the category to which a particular object belongs.
- Regression: Predicting a continuous-valued attribute associated with an object.
- Clustering: Automatic grouping of similar objects into sets.
- Dimension Reduction: Reducing the number of random variables under consideration.
- Model selection: Comparing, validating and choosing parameters and models.
- Pre-processing: Feature extraction and normalization in order to transform input data for use with machine learning algorithm.

**NumPy**

NumPy is the fundamental package for scientific computing with Python. It provides a high-performance multidimensional array object, and tools for working with these arrays. It contains among other things:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities.

**Pandas**

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

**Performance evaluation**

This section discusses about the performance analysis of Naïve Bayes and Logistic Regression on Big data. The methods were employed with the collected samples of datasets containing 1 lakh samples of twitter information. The metrics used here for evaluating the performance are accuracy, precision, recall and F1-score.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$\text{F1 Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

where, $tp$ is true positive which correctly predicted positive values. $tn$ is true negative which correctly predicted negative values. $fp$ is false positive which is falsely predicted positive class and $fn$ is falsely predicted negative class.

The metric representation is demonstrated in the table below:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Accuracy</th>
<th>Pre-cision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>73%</td>
<td>.75</td>
<td>.74</td>
<td>.73</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>69%</td>
<td>.69</td>
<td>.69</td>
<td>.69</td>
</tr>
</tbody>
</table>

The accuracy obtained with Naïve Bayes classification technique on Big data is 73% whereas Logistic Regression produces 69% accuracy. The precision and recall measures
obtained by the Naïve Bayes method are 0.75 and 0.74 whereas Logistic Regression produced 0.69 precision and 0.69 recall value. The F1 score obtained by the Naïve Bayes method is 0.73 whereas Logistic Regression produces 0.69 F1 score value. Thus it is evident from the table that Naïve Bayes Classifier yielded more classification accuracy than Logistic Regression classifier.

CONCLUSION

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analyzing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. The classification accuracies of Naive Bayes and Logistic Regression on the twitter data is compared and the result shows that Naïve Bayes classifier yielded more classification accuracy than Logistic Regression classifier.

REFERENCES


