International Conference
AUTOMATICS AND INFORMATICS'2017
4-6 October 2017, Sofia, Bulgaria

OCEAN BIG DATA PROCESSING IN IoT ECOSYSTEM

D. Ivanova[1], S. Zahov[2]

[1]Technical University of Sofia, Faculty of Applied Mathematics and Informatics, bul. Kl. Ohridski 8, bl.2, office 2541, tel. +359893690007, e-mail: d_ivanova@tu-sofia.bg

[2]Technical University of Sofia, Faculty of German Engineering Education and Industrial Management, bul. Kl. Ohridski 8, bl.10, e-mail: stefan.zahov@fdiba.tu-sofia.bg

Abstract: This paper presents different methods of ocean data collection in the IoT ecosystem. Most of the ocean big data are connecting to sea surface temperature, subsurface temperature, water flows, air mass movement and their interaction at ocean-atmosphere level, sea level, sea ice concentration, and topography of the ocean floor, meteorological conditions and their influence on the ocean surface. All these characteristics of Ocean data are from significant importance with respect to the climate change and its influence on human life. In this paper, the conceptual model for big data analytics of ocean data based on machine learning will be proposed. The experimental framework is based on Apache Spark environment and python3 programming language optimized for big data processing is utilized. The experiments using the linear regression and support vector machine algorithms are conducted. Finally, the result analyses are presented.

Key words: IoT, Ocean Big Data, Machine Learning, Linear Regression, Support Vector Machines, Result Analysis

INTRODUCTION

The oceans constantly operate sensors that constantly collect data about currents, temperature, and the movement of water masses. This data is sent back to scientists and environmental researchers. This information, collected by the sensor networks, helps to address the climate. The sensor network that collected the ocean big data is known as Ocean Internet of Things (OIoT). In the oceanology a broad range of different data types existed. Most of the data collected from the research is related to the physic-geological characteristics of the ocean. Some of them are sea surface temperature (SST), subsurface temperature, water flows (direction, circulation, velocity), air mass movement and their interaction at ocean-atmosphere level, sea level, sea ice concentration, and topography of the ocean floor, meteorological conditions and their influence on the ocean surface. Most of the oceanographic data is collected using two methods: in-situ and remote sensing. Most of the oceanographic data analyses are related to the sea surface, as most radar and other tools used in situ methods, as well as satellite measurements, relate to the surface. Most of the data is analysed by simple methods (through analysed grids), but there are also those that are monitored in real time. They need to be analysed using previous models of data assimilation. [1]

The satellites are one of the most important and used tools for Remote Sensing Research. They receive processed information from ship radars and buoys. But there are other ways to gather information about the ocean surface. The instruments of remote sensing are radiometers, scatterometers, and altimeters. The radiometers measure the sea surface temperatures SSTs. The scatterometers measure the wave disturbances and wind speed and directions. The altimeters measure the ocean surface deformation, "used to estimate sea surface slopes and ocean currents". [2] The information gathered from the Satellites is one of the most important data for oceanographic research.

The in-situ method which is from ships and buoys does not provide enough information for complete analysis of the spatial or temporal data resolution over the enormous ocean districts that cover more than 70% of the planet. Carefully calibrated and adjusted the satellite information, sometimes combined with the in situ observations as a data processing procedure, provides the best degree of global ocean conditions. There is a different level of quality of the useful in situ ocean observation, depending on the sources. One of the best ways (sources, methods) to gather information are scientific research programs, by instrumented buoys, by ships specifically designed to collect environmental data and by coastal or island stations with nearly the same way of functioning as the standard land stations. With lower quality but still important are the frequently collected data by merchant ships through their routes and by fishing fleet vessels during their trade operations. Different scientific research programs collect the widest variety of in-situ data. These types of data are divided into several categories: sea surface data (salinity, SST, wave height, and wave direction), near surface meteorological conditions (air temperature, wind speed, wind direction, and cloudiness), subsurface sea water characteristics (dissolved gases, anthropogenic tracers, and ocean currents). Another large scientific area that collects the ocean big data is the data from high-performance computer simulations. [3]

In this paper, the conceptual model for ocean big data analytics based on machine learning is proposed. The used ocean big data for experiments is collected in the OIoT by using different collection methods.

There are different methods for collection ocean big data through buoys, vessels (ships), high-frequency radars, satellites and data from High Performance Computing (HPC).

Some research programs work through surface drifting buoys (drifting or moored). The location of those buoys is monitored by satellite. This is helpful for the ocean circulation and some geophysical variables. Those buoys are very effective when gathering meteorological and oceanographic data in remote ocean areas. Some buoys are located below the ocean surface and they are followed acoustically or they are tracked by satellite, when they rise to the surface. The data that comes from this buoy are useful in exploring underwater currents, ocean stream, and sea water features. Buoys use accelerometers and inclinometers used to measure wave acceleration and direction.

The source data from these buoys is processed by on-board computer which using statistical analysis generates new data that they later send to coasts stations of continents or native islands. This data (for waves) includes frequency of wave energy, wave height (sometimes the difference between the highest and the lowest), averaged wavelength, and time-out and wavelength stages (on every 20 min). [4]

Another method of collecting data besides buoys is also carried out by ocean vessels. The data collected by commercial and fishing vessels is crucial for the information on the ocean surface. In the 19th century, this was a popular method for collecting meteorological forecast data via the ocean. Modern vessels use automated systems where the multitude of large data is collected in the form of digits and transmitted by satellite to land-based data collection agencies. Input data is derived from wind speed and wind direction, barometric pressure, and air temperature. They are useful for the meteorological forecasts. [5]

Another method for gathering the ocean data is through high-frequency radars, which are set on the shore and scan 24 hours of surface currents in a fairly wide range of up to 200 km. The radar emits electromagnetic waves in the frequency range of 5 to 25 megahertz. These waves spread over the surface of the ocean surface. When the electromagnetic wave encounters an ocean wave (moving to the shore or vice versa), which is half the wavelength, the signal is reflected back to the receiver. Then the frequency and the speed of the returned signal are evaluated. Scientists can thus estimate the size of the ocean wave through the return signal. [6]

Gathering the ocean data through a satellite is the one of the most import method for collecting ocean data. Using this method many ocean factors can be traced, not only to the surface but also to the water. They serve as a precursor to storms, tsunamis and other meteorological consequences. One of the satellites that give accurate results is NASA and is called Toppex/Posledon. This satellite allows a continuous weekly observation of the ocean surface. It also helps to explore the implications of oceans and climatic conditions. The data from this satellite helped to predict study and observe the currents of oceans, prepare routes on ships, help manage fisheries, explore the geochemical and animal world of the ocean. [7]

Finally, HPC in oceanology is the scientific area that generates a large amount of data that is important to be further analysed and visualized.

In this paper, the conceptual model for ocean big data analytics based on machine learning will be proposed.

CONCEPTUAL MODEL FOR OCEAN BIG DATA ANALYTICS

The conceptual model used the data collected in OIoT ecosystem. The ocean data is processed based on machine learning and finally the ocean data is visualised for better result analyses from the oceanology experts, Fig. 1.

In this paper, the results of applying the proposed conceptual model by using two machine learning algorithms Linear Regression and Support Vector Machine (SVM) will be presented and discussed.

Linear regression is the most widely used of all statistical techniques for study of linear, additive relationships between variables and it is commonly used technique for the predictive analysis [8].

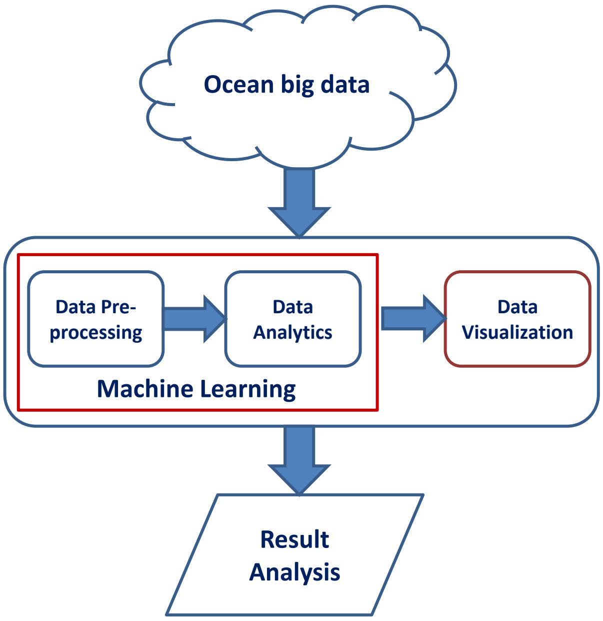


Figure 1: Conceptual Model for Ocean Big Data Analytics

SVM are a useful technique for data classification and regression. They belong to a family of generalized linear classification. A classification task with SMV usually involves separating data into training and testing sets. Each instance in the training set contains one target value and several attributes (features). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes [9].

EXPERIMENTAL FRAMEWORK

The experimental framework is based on Apache Spark environment that allows streaming and real-time big data analyses. For implementation of the framework, python3 programming language optimized for big data processing is utilized. For the case of ocean big data analytics, the NOAA database is used. NOAA is a federal agency in the structure of the US Department of Commerce. It coordinates various types of meteorological and geodetic studies and forecasts in the United States and their possessions, studying the atmosphere and the World Ocean, and also warns the population about possible natural disasters and catastrophes. [10]

EXPERIMENTAL RESULTS AND ANALYSIS

The NOAA database is used to conduct the experiment. Experiment data is collected from a Hawaii station named: Honolulu, HI - Station ID: 1612340. With 5 individual sensors, the station collects data on water level, wind speed, air temperature, water temperature and pressure Water. The sensor through which we collect the water level data to be analyzed is activated in 6 minutes. So it writes new data over 6 minutes during a time from 00:00 to 13:12h in a CSV file. Thus, all recorded data is 133. The new information that the sensor counts is divided by 9 attributes, three of which are essential for the data analysis*: time, water level (in meters) and sigma value.* The importance of the sigma value serves for a better understanding of the wave level. To better understand what the significant wave height is, we need to look at the statistics and allocation theory.

The most common distribution is the normal distribution, where participation also takes the sigma. This allocation is useful in describing data where most elements in a class are grouped close to an average, with an equal number of elements being greater or less than that average.

What the software first does is to plot the water height and the relative sigma over same time in the form of a curve, Fig.3. From the same database, we download a CSV file with the water temperature values for the same time and date. In Fig.4, it can be seen the distribution of the water temperature values on the same day measured at the same station and detected at the same time interval. Finally, the distribution of points by the ratio of wave height and sigma value, which we are going to examine and analyze by two machine learning algorithm, is done.

For the big data analytics of collected data, two machine learning algorithms will be used: Linear Regression and SVM. For the first experiment, our data analysis begins with the study of the relationship between water level and its sigma value. We use one of the most common machine learning algorithms, which is called linear regression. Linear regression is a statistical method for constructing a (possibly) acceptable linear relationship between a set of independent variables x1, x2, ..., xm and the dependent variable y (control magnitude). We build a linear mathematical model that can help estimate the state of y data at different x data. For the equation of the straight line we use the following formula for Cartesian linear equation: *(y = a + bx)*, where *a* is the point of intersection of the abscissa X, and *b* is the coefficient of inclination. They are calculated using the following formulas known in the statistics:

$$a = \frac{\left(\sum_{j=1}^{n} y_j\right)\left(\sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} x_j y_j\right)}{n\left(\sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j\right)^2} \qquad (1)$$

$$b = \frac{n\left(\sum_{j=1}^{n} x_j y_j\right) - \left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)}{n\left(\sum_{j=1}^{n} x_j^2\right) - \left(\sum_{j=1}^{n} x_j\right)^2} \qquad (2)$$

In these formulas, Y and X denote two sets of input data and have no common with x and y in the general formula for the Cartesian linear equation. Linear regression finds the relationship between these two sets of input data.

We use these formulas by writing them in the form of a Python3 program and taking the datasets from NOAA, with the set X being all 133 values of the wave heights that are registered from the sensor, and the set Y are all corresponding values of the sigma. Thus, we build regression rights that pass through the points and divide them as much as possible. From the compilation of the code we get the equation of the line, Fig.5. We get the point of intersection and the point of inclination. This equation for given arbitrary values of the new variable *x* builds the graphs. Thus, a linear mathematical model is constructed to help estimate the state of a y for different x data. The result of the experiment we made about linear regression allows us to make a correct forecast for incoming set X with input data. In our case - assisting for a correct forecast for input data - water level and its value of the sigma.

Our second experiments performed SVM algorithms. SVM algorithms are a new promising method for classifying and regrouping data. The main task of these algorithms is to depict points (experiments) in the plane that must be classified by a particular line in such a way as to make an assessment that is as accurate as possible. In a set of training examples, each identified as belonging to one of two categories, the SVMs training algorithm creates a model that predicts when a new example falls into the first category or second. This is the so-called binary classification when, for example, the machine is trained to classify newcomers in 2 classes.

There are two basic concepts for SVMs: the large margin and the kernel functions. There are three types of classification functions: *linear* (with simple equation of straight), *polynomial* (with a simple equation of a curve of degree d) and *Gaussian - rbf* (with a gausse curve equation at a value of a certain sigma (standard deviation value) of the distribution between 0 and 1).

In our experiment, we write to Python3 a SVM program using the *svm library* and the *svm.SVR* method to create a classifier that breaks down into 2 classes (in dependence of wave height) the workout set by making a valid appendix assessment. The classes are determined according to the height of the wave, with the first class containing waves with a height of 0.600 to 0.900 m and in the other class - waves with heights from 0.02 to 0.500 meters. Thus, according to these two classes, we train a machine with input vectors to give a precise classification. The validation data is passed through the machine (program), which is classified with one of the both given classes. We use 5 types of regressions - *linear (by means of spikes with bending and crossing odds), polynomial (by constructing a square and cubic curve), and Gaussian - rbf (by constructing a Gaussian curve with a sigma value of 0.1 and 0.01).* From the experiment, we see in table 1 that the closest results yield linear kernel of SVM.

Table 1: Experimental Results of Testing Inputs Data using different SVM Kernels

Kernel	Linear	Polynomial d = 2	Polynomial d = 3	Rbf σ=0.1	Rbf σ=0.01
Precision	98%	54%	50%	53%	48%

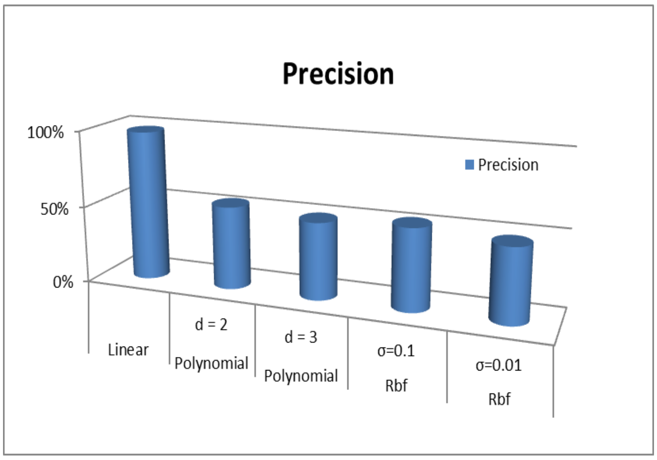


Figure 2: Precision Results of SVM Classification

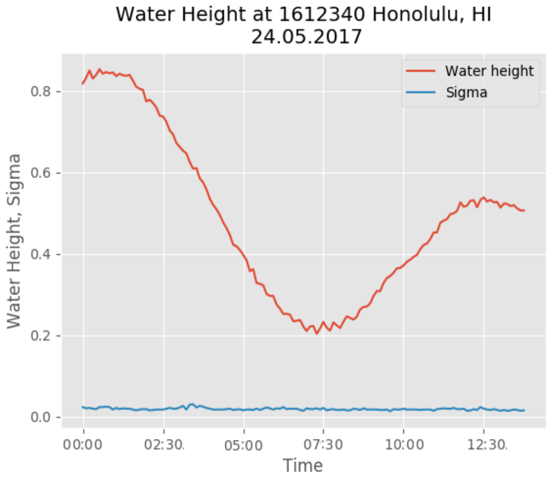


Figure 3: Water Height and the Relative Sigma

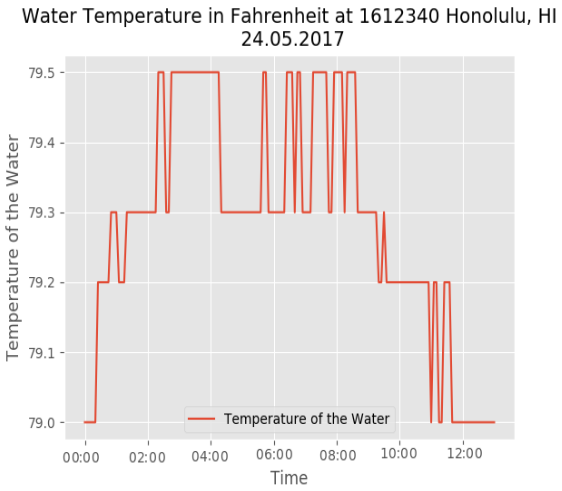
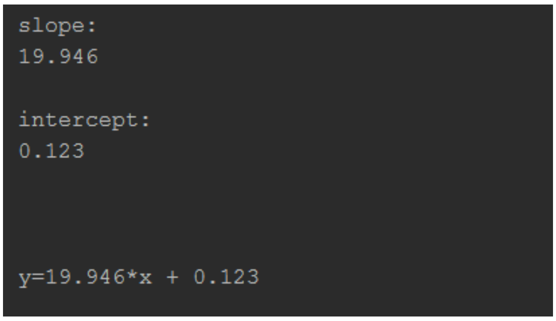


Figure 4: Distribution of the Water Temperature



```
slope:
19.946

intercept:
0.123



y=19.946*x + 0.123
```

Figure 5: Linear Regression Equation Result

We have good computational results of implementing linear kernel of SVM algorithm that achieve the result of 98% precision, because the distribution of data is even. The second most accurate classification is the second degree (d=2) polynomial kernel: $y = (a + bx)^d$ with precision = 54%, Tab.1. Although it is second only to the accuracy of the calculations, the results are not good and are far from an accurate estimate. The reason for this is the graph of the function that is the parabola (curve) that passes, reaches the maximum, and then falls downward, moving away from some vectors. The most distant estimate is obtained with Gaussian regression with a sigma in the equation = 0.01. The reason for this is the curve that closes almost in a circle. At smaller sigma values closes the curve of the equation in a circle / ellipse.

FUTURE WORK

The future work of the authors will be to extend the functionality of SVM program with more classes defining vectors with more attributes of the experimental ocean data and to performed additional experiments with big data sets. The second task of the team will be to realise the proposed conceptual model for ocean big data analytics with more machine learning techniques and algorithms like clustering (KMeans) and Principle Component Analysis (PCA), as well as to provide the comparative analysis of the results with respect to ocean big data processing and analytics.

ACKNOWLEDGMENT

The results reported in this paper are part of a research project № 171ПР0015-11 supported by the Research and Development Sector, Technical University of Sofia.

REFERENCES

1. Silicon Labs Community, Meet the Captain of Oceanic Internet of Things, white paper, 2015.
2. Seneviratne SI, Nicholls N, Easterling D, Goodess CM, Kanae S, Kossin J, Luo Y, Marengo J, McInnes K, Rahimi M, Reichstein M, Sorteberg A, Vera C, Zhang X (2012) Changes in climate extremes and their impacts on the natural physical environment. In: Field CB, Barros V, Stocer TF, Qin D, Dokken DJ, Ebi KL , Mastrandrea MD, Mach KJ, Plattner G-K, Allen SK, Tignor M, Midgley PM (eds) Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of working groups I and II of the Intergovernmental Panel on climate change (IPCC). Cambridge University Press, Cambridge, UK/New York, pp 109–230.
3. Worley, J. Steven, Stern, R. Ilana, An Introduction to Atmospheric and Oceanographic Data, NCAR/TN-404+IA, NCAR TECHNICAL NOTE, August 1994, p.29-40.
4. Banholzer Sandra, Kossin James, Donner Simon, The Impact of Climate Change on Natural Disasters, Springer Science+Business Media Dordrecht 2014, DOI 10.1007/978-94-017-8598-3_2, pp. 21-49.
5. Smith, R. H., and E. Johns (2012). Oceanographic data are collected in many different ways. In: Tropical Connections: South Florida’s Marine Environment, W. L. Kruczynski and P. J. Fletcher (eds.). IAN Press, University of Maryland Center for Environmental Science, Cambridge, Maryland, 66.
6. https://tidesandcurrents.noaa.gov/hfradar/
7. Buis Alan, Hupp Erica, Moreaux Eliane, NASA's Topex/Poseidon Oceanography Mission Ends, 01.05.06
8. http://www.statisticshowto.com/how-to-find-a-linear-regression-equation/
9. L.P. Wang, "Support Vector Machines: Theory and Application," Springer, Berlin, 2005.
10. http://www.ndbc.noaa.gov/cman.php