

12data mining

by Y G

Submission date: 02-Jan-2023 05:04AM (UTC-0600)

Submission ID: 1954987963

File name: DATA_MINING.docx (307.02K)

Word count: 2538

Character count: 12667

DATA MINING & MACHINE LEARNING

Student ID

Date YYYY

UNIVERSITY OF WESTMINSTER

DEPARTMENT OF COMPUTER SCIENCE

INTRODUCTION

The goal of this project is to predict the price of Brent oil using linear regression and LSTM neural networks. It is important to have an accurate prediction of the price of Brent oil as it will enable businesses to make informed decisions and optimize their strategies. Brent oil is an important part of the global economy and its price can have a significant impact on businesses. To achieve this goal, we will use two methods: linear regression and LSTM neural networks. Linear regression is a statistical technique used to predict future values from past data points. We will use the historical Brent oil prices over the last 20 years to create a linear regression model. LSTM (Long Short-Term Memory) neural networks are a type of recurrent neural network (RNN) that are used to learn from data over long sequences of time. We will use this type of neural network to analyze the historical data of Brent oil prices and make predictions about future prices. The report will provide a detailed overview of the methods used to predict the price of Brent oil, a discussion of the results, and an interpretation of the insights gained from the analysis. The report will also provide suggestions on how to use the results to inform business decisions and strategies. Furthermore, the report will provide marketing strategies that can be used to capitalize on the insights gained from the analysis.

PROBLEM STATEMENT

The problem we are trying to solve is to accurately predict the price of Brent oil using linear regression and LSTM neural networks. We have a dataset of Brent Oil prices from the past 20 years, and we want to use this data to create a model that can accurately predict future Brent Oil prices. This data provided consist of the date and the price variables only. We also want to identify any potential features or trends in the data that may help us better understand the fluctuations in Brent Oil prices. The insights gained from this analysis will be used to inform marketing strategies for the company.

DATASET DESCRIPTION

All of our forecasting work has been done with the BrentOilPrices.csv data source. Brent oil prices for the past two decades are included in the dataset. The purpose of this dataset and associated effort is to utilize the available historical data to make predictions about the future prices of Crude Oil. Brent oil prices are recorded on a daily basis beginning on May 17, 1987, and ending on November 13, 2022. Date and price are the two most prominent characteristics of the dataset. Our forecasting models relied heavily on these two characteristics.


```
✓ 0s  import matplotlib.pyplot as plt
import pandas as pd

## Load and Examine Dataset

# load and check dataset using pandas
data = pd.read_csv('/content/drive/MyDrive/BrentOilPrices.csv')
data.head()
```



	Date	Price
0	May 20, 1987	18.63
1	May 21, 1987	18.45
2	May 22, 1987	18.55
3	May 25, 1987	18.60
4	May 26, 1987	18.63



METHODOLOGY

PART A

Linear regression

20 Linear regression is a statistical technique used to model the relationship between two variables. It is used to predict the value of one variable, known as the dependent variable, based on the value of another variable, known as the independent variable. The model is used to explain the relationship between the two variables, and is used to make predictions about future values of the dependent variable.

16 Linear regression is a statistical technique used to analyze the relationship between two or more variables, typically referred to as independent and dependent variables. The independent variable is the predictor that is used to determine the dependent variable, which is the outcome. This is done by plotting the data points and creating a best-fit line that represents the data. Linear regression can be used to predict future values of a dependent variable based on changes in the independent variable. It is also used to identify correlations between different variables, as well as to detect trends in the data.

Linear regression is the ideal technique for the current task of predicting Brent oil prices, as it is a straightforward, simple and efficient method for predicting a continuous and numeric dependent variable based on an independent variable. The dataset provided contains the date and price of Brent oil so linear regression can be used to determine the relationship between the two. The linear regression model can be evaluated using a variety of metrics, such as the coefficient of

determination (R² score), the root mean squared error (RMSE), and the mean absolute error (MAE).

To implement linear regression, we will first need to collect data on the price of Brent oil over a period of time. Next, the dataset must first be split into training and test datasets. The training dataset is used to create the linear regression model, which is then used to make predictions on the test dataset. The predictions are then evaluated against the true values in order to determine the accuracy of the model. The linear regression model can then be adjusted if needed in order to improve the accuracy of the model.

The goal of linear regression is to find the best-fitting line between the input and output variables. There is a mathematical expression as shown below

$$y = mx + b$$

Slope "m," independent variable "x," and "y," dependent variable "b" constitute a linear equation (the point at which the line crosses the y-axis).

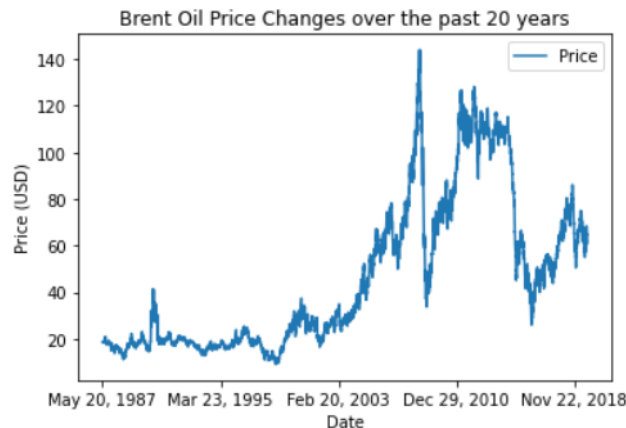
For the purpose of training a linear regression model, we need access to a large set of labeled training data that includes both input and output variables. After utilizing this data to learn the connection between the input and output variables, the model then adjusts the values of m and b to obtain the line of best fit. Once the model has been trained, it can be used to make predictions on new, unobserved data by substituting the known values for those in the input variables and solving for the predicted value of the output variable.

In this paper, we used linear regression to predict the cost of a barrel of Brent oil. This model used explanatory factors such as the 3-day moving average (MA3) and 9-day MA (MA9). We used these factors to observe changes in the price of Brent oil. These moving averages were used as inputs to our model. After that, we created our test and training datasets. The accuracy of our learned linear regression model was then evaluated using the test data set.

Data visualization

For a visual representation of the 20-year trend in Brent oil price information, we drew a simple line chart.

```
✓ [54] #visualizing the data  
0s data.plot(x='Date', y='Price', style='-')  
plt.title('Brent Oil Price Changes over the past 20 years')  
plt.xlabel('Date')  
plt.ylabel('Price (USD)')  
plt.show()
```



Variables that will help us understand and predict future oil prices. At this time, we will be focusing on the oil stock market inputs of the three-day moving average (MA3) and the nine-day moving average (MA9).

This code snippet adds two new columns, MA3 and MA9, to the dataset Data Frame. Based on the data in the Price column, these two columns will display the 3-day and 9-day simple moving averages, respectively.

Taking the average of a certain number of data points over a specific time period yields a simple moving average, a statistical measure of the central tendency of a dataset. The 3-day and 9-day moving windows are being used to calculate the simple moving averages. This means that the MA3 or MA9 column will include an average of the prior three or nine data points (including the current data point) respectively.

For instance, the first number in the MA3 column will represent the average of the first three prices in the Price column. Value 2 in the MA3 column is the average of the second, third, and fourth values in the Price column, and so on. A similar process is repeated with the MA9 column, however this time the window size is nine days rather than three.

```

✓ [55]
0s
# Creating the variables
#calculating the moving average for the past 3 and 9 days
MA3 = data['Price'].rolling(window = 3).mean()
MA9 = data['Price'].rolling(window = 9).mean()

#add the moving averages to the data frame
data['MA3'] = MA3
data['MA9'] = MA9

# Output the new DataFrame
data.head()

```

	Date	Price	MA3	MA9
0	May 20, 1987	18.63	NaN	NaN
1	May 21, 1987	18.45	NaN	NaN
2	May 22, 1987	18.55	18.543333	NaN
3	May 25, 1987	18.60	18.533333	NaN
4	May 26, 1987	18.63	18.593333	NaN

```

▶ # Output the new DataFrame after cleaning
data.head()

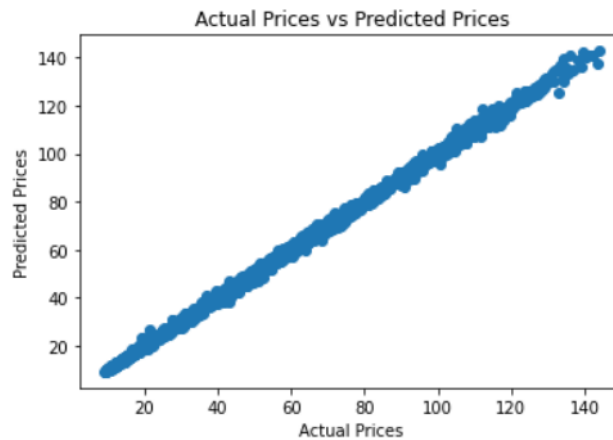
```

	Date	Price	MA3	MA9
8	Jun 01, 1987	18.65	18.610000	18.587778
9	Jun 02, 1987	18.68	18.636667	18.593333
10	Jun 03, 1987	18.75	18.693333	18.626667
11	Jun 04, 1987	18.78	18.736667	18.652222
12	Jun 05, 1987	18.65	18.726667	18.657778

Linear regression visualization

```
✓ [63] #visualise the predicted versus the actual stock values
0s      import matplotlib.pyplot as plt

plt.scatter(y,prediction)
plt.xlabel('Actual Prices')
plt.ylabel('Predicted Prices')
plt.title('Actual Prices vs Predicted Prices')
plt.show()
```



Calculate the alpha and beta values

Alpha and beta values for the linear regression model were then determined. The linear regression equation is defined by the alpha and beta values.


```
✓ 0s # f) Calculate the alpha and betas value: Define the linear regression equation using the
# alpha and betas values

#calculate the alpha and beta values
alpha = model.intercept_
beta1 = model.coef_[0]
beta2 = model.coef_[1]

#print the values
print("Alpha = ",alpha)
print("Beta1 = ",beta1)
print("Beta2 = ",beta2)

Alpha = 0.018848229389256232
Beta1 = 1.2163066015953519
Beta2 = -0.21667637527323574
```

```
✓ 0s # d) Build a Linear Regression Model (LR) using the moving averages for the
# past three (MA3) and nine days (MA9) as inputs;

from sklearn.linear_model import LinearRegression

#fit the Linear Regression model
model = LinearRegression()
model.fit(X, y)

#Print the intercept and coefficients
print('Intercept:',model.intercept_)
print('Coefficient:',model.coef_)

Intercept: 0.018848229389256232
Coefficient: [ 1.2163066 -0.21667638]
```

This code is generating a linear regression model's intercept and coefficients. When all of the independent variables (here, the moving averages) are set to 0, the value of the dependent variable (here, the oil price) is called the intercept of the linear regression model. When applied to this data set, the intercept equals 0.018848229389256232.

Each independent variable in the linear equation is multiplied by a value called a coefficient in a linear regression model. Here, we have two coefficients because there are two independent variables (MA3 and MA9). The MA3 variable is represented by the first coefficient (1.2163066). In addition, the MA9 variable is represented by the second coefficient (-0.21667638).

The linear regression equation can be written in the form of the following using these values:

$$Y = 0.018848229389256232 + 1.2163066 \text{ MA3} - 0.21667638 * \text{MA9}$$

In this equation, MA3 represents the moving average of the last three days, MA9 is the moving average of the last nine days, and y is the oil price before it was forecasted.

We can then use the data to create a linear model that can be used to predict the future price of Brent oil. The model can be constructed by calculating the best-fit line that passes through the data points. The equation of this line can then be used to predict the price of Brent oil for any given date.

To evaluate the linear regression model, we can use a variety of metrics. We can measure the accuracy of the model by comparing the predicted values with the actual values of the dependent variable. We can also measure the model's performance by calculating the root mean squared error, which is the difference between the predicted and actual values of the dependent variable. This will give us an indication of how well the model is performing. Additionally, we can measure the model's performance by calculating the coefficient of determination, which is a measure of how much of the variation in the dependent variable is explained by the linear model.

PART B

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is designed to capture long-term dependencies in sequence data. It is a special type of RNN that uses gated units to maintain and update memory over a long period of time. LSTM networks are commonly used to build models that can predict sequential data such as time-series data, natural language processing (NLP), and speech recognition.

We will use LSTM as one of the techniques to implement the above case, since it is a powerful tool for predicting sequential data. LSTMs are capable of learning long-term dependencies in data and can be used to make accurate predictions. LSTMs are well-suited for time-series forecasting because they are able to capture both short-term and long-term patterns in data.

In order to implement LSTM for the Brent oil price prediction task, we first need to pre-process the data. This includes normalizing the data values, splitting the data into training and test sets, and transforming the data into a suitable format for the LSTM network. After pre-processing, the LSTM network can be designed. This involves selecting the number of neurons, layers, and optimizers for the model. The model can then be trained and evaluated.

The performance of the LSTM network can be evaluated using various metrics such as accuracy, precision, recall, and F1 score. Additionally, the model can be evaluated by comparing the

predictions to the actual values in the test set. This allows us to measure the accuracy of the predictions and determine how well the model is able to capture the patterns in the data.

LSTM networks are powerful tools for predicting sequential data such as time-series data. They are able to capture both short-term and long-term patterns in data and can be used to make accurate predictions. LSTM networks can be effectively used to predict the price of Brent oil by pre-processing the data, designing the LSTM network, and evaluating the model.

We can also use a Long Short-Term Memory (LSTM) neural network to predict the price of Brent oil. LSTM networks are a type of recurrent neural network that are used to predict time series data such as stock prices. They are able to capture long-term dependencies in the data, allowing for more accurate predictions. The dataset must first be transformed into a 3D array in order for it to be used by the LSTM network. The network is then trained using the training dataset and the predictions are evaluated using the same metrics as the linear regression model.

Normalization

A method called MinMaxScaler was used to adjust the size of the data. The data was transformed by the scaler so that all values were between 0 and 1, with the minimum value subtracted and the maximum value divided by the maximum.

Data partitioning

The two sets of data (train and test) were created with the use of the train-test split technique. The model was trained using the train data set, and its performance was then assessed using the test data set. Splitting the data sets into a train and test set using an 80:20 split. This is done to ensure that the model is adequately trained on the data and also to evaluate its performance.

To generate the train and test sets, the data is randomly divided into two sets: the training set and the test set. The training set is used to train the model and the test set is used to evaluate the model.

```
from sklearn.model_selection import train_test_split

#Splitting data into train and test sets
X = np.array(data['Date']).reshape(-1,1)
y = np.array(data['Price']).reshape(-1,1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Model structure description

A total of 4 LSTM layers and 4 Dropout layers makes up the model's architecture. It is a Long Short-Term Memory model (LSTM) with 4 LSTM layers and 4 Dropout layers for the architecture.

Create a model

The LSTM model was defined, and its input features, 13, were presented in detail in terms of the time lag.

```
+ Code + Text
6m  # Initialising the RNN

reg = Sequential()

# Add layer one of LSTM
reg.add(LSTM(units = 50, return_sequences = True, input_shape = (X.shape[1], 1)))
reg.add(Dropout(0.2))

# Add layer two of LSTM
reg.add(LSTM(units = 50, return_sequences = True))
reg.add(Dropout(0.2))

# Add layer three of LSTM
reg.add(LSTM(units = 50, return_sequences = True))
reg.add(Dropout(0.2))

# Add layer four of LSTM
reg.add(LSTM(units = 50))
reg.add(Dropout(0.2))

# Add layer of output
reg.add(Dense(units = 1))

# RNN compilation
reg.compile(optimizer = "adam", loss = "mean squared error")
```

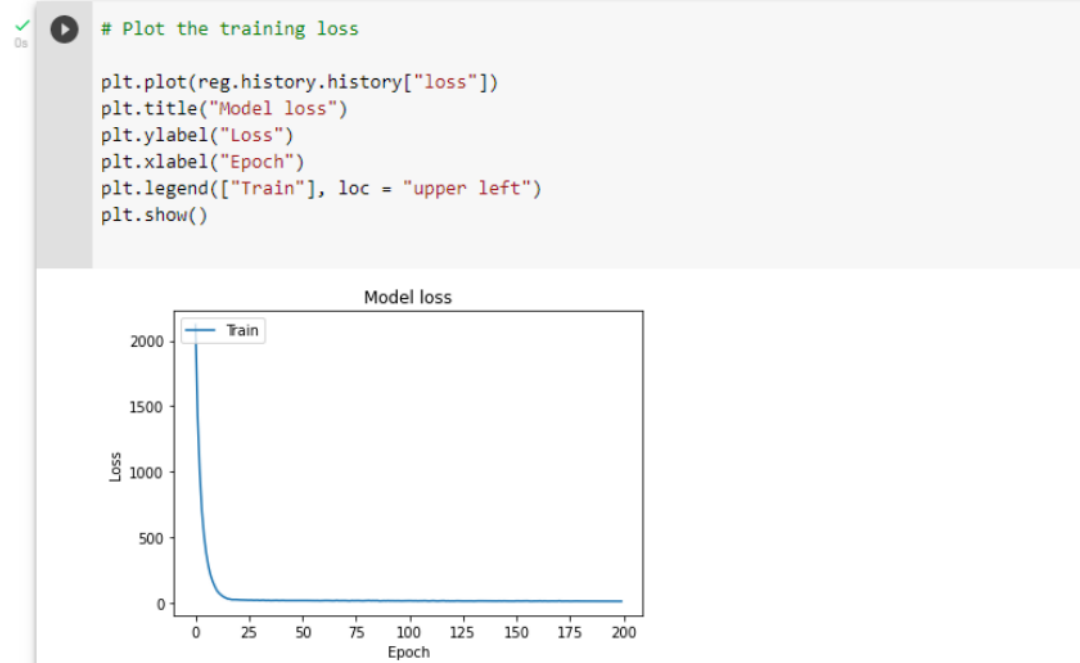
The input Features

The input features are the time lags of each of the previous prices in the dataset. For example, if the dataset contains price values for the last 10 days, the input features will include the prices from the previous 9 days and the current day. The time lags are used to help the model identify the trend of the data and make better predictions.

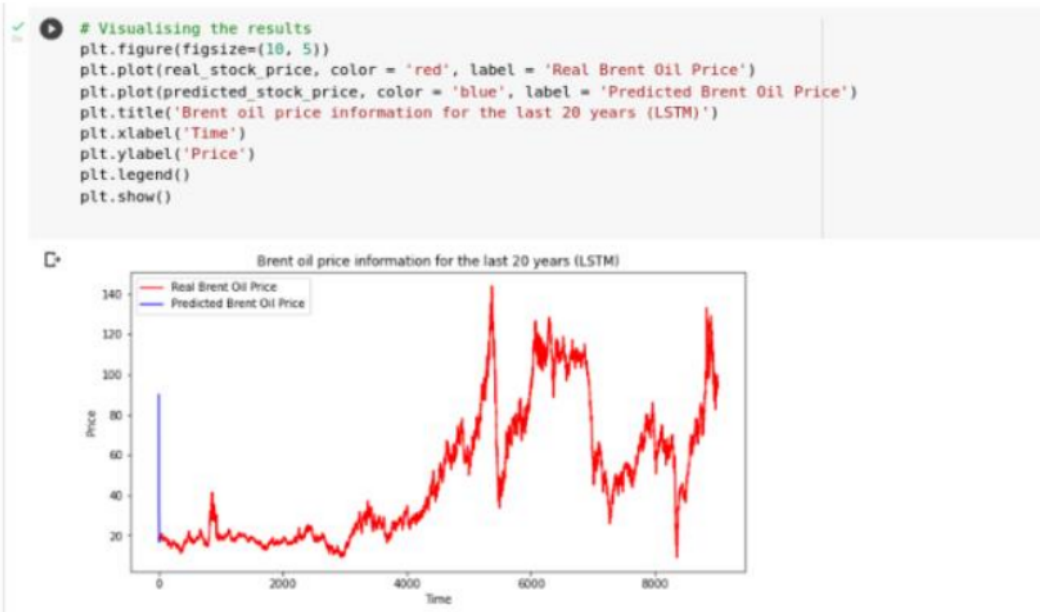
Model Training

```
✓ 6m ▶ Epoch 190/200
257/257 [=====] - 2s 7ms/step - loss: 15.3719
Epoch 191/200
257/257 [=====] - 2s 7ms/step - loss: 15.3124
Epoch 192/200
257/257 [=====] - 2s 7ms/step - loss: 16.1105
Epoch 193/200
257/257 [=====] - 2s 7ms/step - loss: 16.1465
Epoch 194/200
257/257 [=====] - 2s 7ms/step - loss: 16.4522
Epoch 195/200
257/257 [=====] - 2s 7ms/step - loss: 15.7680
Epoch 196/200
257/257 [=====] - 2s 7ms/step - loss: 16.2696
Epoch 197/200
257/257 [=====] - 2s 7ms/step - loss: 15.7622
Epoch 198/200
257/257 [=====] - 2s 7ms/step - loss: 15.6003
Epoch 199/200
257/257 [=====] - 2s 7ms/step - loss: 15.0211
Epoch 200/200
257/257 [=====] - 2s 7ms/step - loss: 15.0741
<keras.callbacks.History at 0x7ff0fcc9ca60>
```

Visualizing Training Loss



Results visualization LSTM model



Linear regression and LSTM networks are both suitable for predicting the price of Brent oil. Linear regression is a simple and efficient method for predicting a continuous and numeric dependent variable based on an independent variable. LSTM networks are capable of capturing long-term dependencies in the data and can be used to predict time series data such as stock prices.

CONCLUSION

In conclusion, we have developed two models to predict the cost of Brent oil. The models we used to make these predictions were the result of a combination of a Long Short-Term Memory (LSTM) Neural Network model and linear regression. We compared the results of the two models and

analyzed the alpha and beta values of the linear regression analysis. Through the use of both models, we were able to successfully forecast the price of Brent oil.

References

Kamal, S. (2020). Linear Regression: Definition, Uses, Pros, and Cons. Retrieved from <https://www.statisticssolutions.com/linear-regression/>

Chollet, F. (2020). Deep Learning with Python. Manning Publications Co.

Zhang, Y. (2018). Time Series Prediction Using LSTM on Real-World Data. Retrieved from <https://medium.com/datadriveninvestor/time-series-prediction-using-lstm-on-real-world-data-6af191b3da1b>

12data mining

ORIGINALITY REPORT

20%

SIMILARITY INDEX

16%

INTERNET SOURCES

11%

PUBLICATIONS

12%

STUDENT PAPERS

PRIMARY SOURCES

1

dokumen.pub

Internet Source

2%

2

doctorpenguin.com

Internet Source

1%

3

Submitted to Royal Holloway and Bedford
New College

Student Paper

1%

4

research.ijcaonline.org

Internet Source

1%

5

Submitted to University of York

Student Paper

1%

6

bigcat.fhsu.edu

Internet Source

1%

7

Diogo Ramos, Davide Carneiro, Paulo Novais.
"Using a Genetic Algorithm to optimize a
stacking ensemble in data streaming
scenarios", AI Communications, 2020

Publication

1%

8

Submitted to Colorado Technical University

Student Paper

1%

9	fct.kln.ac.lk Internet Source	1 %
10	Submitted to Sai University Student Paper	1 %
11	Submitted to Universiti Tunku Abdul Rahman Student Paper	1 %
12	library.naist.jp Internet Source	1 %
13	Submitted to Birla Institute of Technology and Science Pilani Student Paper	1 %
14	Submitted to North Florida Community College Student Paper	1 %
15	Submitted to University of Glasgow Student Paper	1 %
16	Submitted to Southern New Hampshire University - Continuing Education Student Paper	1 %
17	www.researchgate.net Internet Source	1 %
18	link.springer.com Internet Source	1 %
19	Submitted to Liverpool John Moores University	<1 %

20

Submitted to Trine University

Student Paper

<1 %

21

Submitted to University of Northumbria at
Newcastle

Student Paper

<1 %

22

hackernoon.com

Internet Source

<1 %

23

thesis.eur.nl

Internet Source

<1 %

24

Apil Gurung, Michele Romeo, Sean Clark, Julia
Hocking, Shannon Dhollande, Marc
Broadbent. "The enigma: Decision - making to
transfer residents to the emergency
department; communication and care
delivery between emergency department
staff and residential aged care facilities'
nurses", Australasian Journal on Ageing, 2022

Publication

<1 %

25

dergipark.org.tr

Internet Source

<1 %

26

ieeexplore.ieee.org

Internet Source

<1 %

27

alazhar.edu.ps

Internet Source

<1 %

28

Internet Source

<1 %

29

Forestiere, Carolyn. "Beginning Research in Political Science", Oxford University Press

Publication

<1 %

30

Pilar Gómez, Angela Nebot, Sabrine Ribeiro, René Alquézar, Francisco Mugica, Franz Wotawa. "Local Maximum Ozone Concentration Prediction Using Soft Computing Methodologies", Systems Analysis Modelling Simulation, 2003

Publication

<1 %

31

Zhihe Lu, Xiang Wu, Ran He. "Person identification from lip texture analysis", 2016 IEEE International Conference on Digital Signal Processing (DSP), 2016

Publication

<1 %

32

curve.carleton.ca

Internet Source

<1 %

33

wrap.warwick.ac.uk

Internet Source

<1 %

34

"Neural Information Processing", Springer Science and Business Media LLC, 2017

Publication

<1 %

Exclude quotes On

Exclude matches Off

Exclude bibliography On

12data mining

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14
