A Hadoop based Weather Prediction Model for Classification of Weather Data

A.K.Pandey
Department of Information Technology,
KIET Group of Institution, Muradnagar, Ghaziabad
ak.pandey@kiet.edu

C. P Agrawal
Department of Computer Science and Applications, M.C.N.U.J.C., Bhopal, India
agrawalcp@yahoo.com

Meena Agrawal
Department of Energy, M.A.N.I.T, Bhopal, India
oshomeena@gmail.com

Abstract-- *Big Data is spreading vastly in the industry. Most of the industries want to have the records of not only the work they do but also are eager to know the taste of the consumer. Big Data is becoming relative to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer. In this work, application of big data is investigated in the field of weather prediction. The weather prediction data is generated from various sources such as radar, ships, ground observation etc. It contains certain useful and useless information for prediction of weather data and in the form of unstructured data. Further, in this work, Hadoop framework is applied to process this unstructured data. The word count algorithm is being used to find the overall condition of that day. Further fuzzy logic (FL) and artificial neural network fuzzy interface system (ANFIS) methods are investigated for accurate prediction of weather data on the basis of mean square error. Experimental results show that the ANFIS method gives more accurate results in comparison to other methods being compared.*

Keywords: Weather prediction, ANFIS, FL and ANN

1. INTRODUCTION

In the present era, weather forecasting is an essential process that can affect several areas such as agriculture, farming, sea selling etc. and also for saving the life of living beings from climate hazard, earthquake and many more. Weather forecasting can be understood as interpretation of atmospheric data which includes temperature, rainfall, wind speed, wind direction and humidity. These conditions can be changed dynamically. Large numbers of tools are available to forecast the weather data, but the amount of data generate for weather forecasting is of high volume and unstructured. Hence, to predict the weather on the behalf of the weather data is not an easy task and it involves large number of parameters that can rapidly change as per atmospheric conditions. To prevent the climatic hazard in future due to weather, numbers of meteorological departments are working in this direction through sharing of information. Weather forecasting can be viewed as significantly challenging problem and can be required the latest technology equipment and technology for accurate prediction of future prediction. It consists of two factors i.e.

human activities and technological advances [1]. For accurate prediction, the various researchers have identified

meteorological characteristics by applying different methods. It is observed that few of method predict the weather more accurate in comparison to others [2].The weather forecasting basically measures the change occurred in present state of atmosphere. The present condition of weather is obtained from the observation like ground observation, from sea observation and the observation of air data through radars. Numbers of device are used to collect this information such as ship, radar, aircraft, satellites and radio sounds. After collecting the desired information, it can be processed and analyze to find the different patterns. In this work, large numbers of high end computers are used to process the data and tried to find the significant and effective piece of information. Present time, computers are widely used to for weather forecasting, called numerical weather prediction. In this process, some models are developed to measure the atmospheric changes. Basically, these models consist of numerical equation that can observe the changes in meteorological characteristics such as atmospheric temperature, pressure, and moisture with respect to time. These equations are solved using the computer program and computer program compute this equation quickly with rapid change in data. This procedure is repeated again and again to forecast the weather and output can be showed in terms prognostic chart.

In past few decades, numbers of weather forecast models are developed and work on image acquisition process. In

978-1-5090-3239-6/17/$31.00©2017 IEEE

1996, RAMS model has been developed for better forecasting of the weather of Atlanta city of USA [3]. It is a cluster based meteorological system for forecasting of the weather. The different dynamic models are being designed for prediction of the weather [4]. A Knowledge Based System for WeatherInformation Processing and Forecasting has also been reported in [5]. This model consists of five components such as image acquisition, image processing and enhancement, feature extractionweather knowledge base and weather inference engine. This system works on satellite image and the meteorological characteristics. A weather forecast model based on soft computing technique is also presented in [6]. This model also works on images and processes the images for obtaining the actual data and forecast the data based on past records and history. To visualize the relevant information to its users, weather research forecast portal has developed a self-configurable weather research forecast model for configuring and scheduling specific weather forecast. It generates weather visualizations relevant to its audience [7]. An association rules based forecasting model is also presented in literature [8]. For effective management of air traffic, a traffic flow management has been designed [9]. It is also found that ANN can be applied for predictionandclassificationof thunderstormwith good accuracy rate [12].

2. HADOOP BASED WEATHER FORECASTING MODEL

In this work, a Hadoop based weather forecasting model is proposed for efficient processing and prediction of weather data. The complete process has been completed in following steps:

Step-1: In order to predict the weather data, firstly, the data is collected from the sourcehttps://www.wunderground.com/history/airport/VIDP/2016/2/27/DailyHistory.html?HideSpecis=1

Step-2: The collected data is preprocessed by using word count algorithm of HADOOP.

Step-3: Data Set has been described and its classification attributes are identified.

Step-4: The prediction has been done by using the ANFIS and Fuzzy Logic.

Step-5: Compare the result.

2.1 Data Acquisition

. To collect the desired data, the html of this website is parsed and data stored in .csv form. Beautiful soup is used to parse the html and fetch the data of a day. It is a Python library for pulling data out of HTML and XML files. It works with parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. Fig. 1 shows the snapshot of the acquired data The following syntax is used to fetch the data

```
$ apt-get install python-bs4
$ easy_install beautifulsoup4
$ pip install beautifulsoup4
$ apt-get install python-lxml
$ easy_installlxml
$ pip install lxml
```

In this work, ten years weather data is collected. In order to get the complete data, a script is written using Python and BeautifulSoup. A set of date is generated using the datetime. Datetime is a library that can be easily imported and used. From Urlparse, the urlsplit and urlunsplit has been used that is used to generate url of a particular day and then it is being parsed using BeautifulSoup. The result retreived is stored in CSV(Comma Seperated Values) format.Fig. 2 shows the shows the snapshot of the script that are used to retrieve the dataset. Fig. 3 shows the screenshot of the script which are used to convert the CSV file to text file.

2.2 Data Pre-processing

In HADOOP for data processing, word count algorithm is applied on txt format. For this purpose, the data files of each day are converted to txt form using a bash script. This script can be run on the complete folder of the data and it

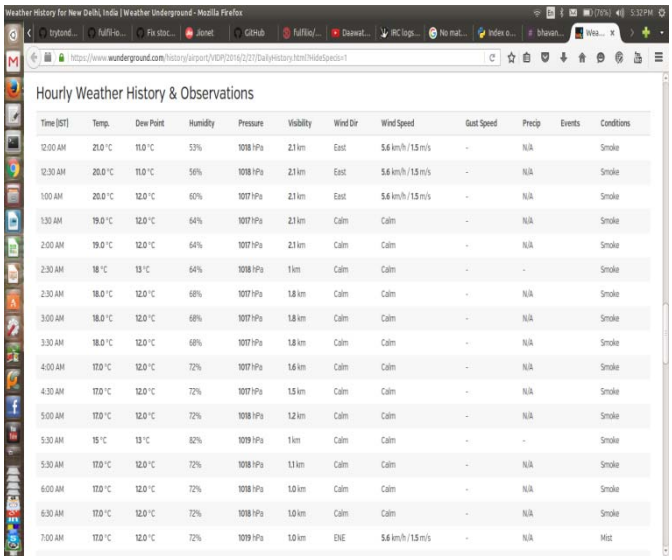
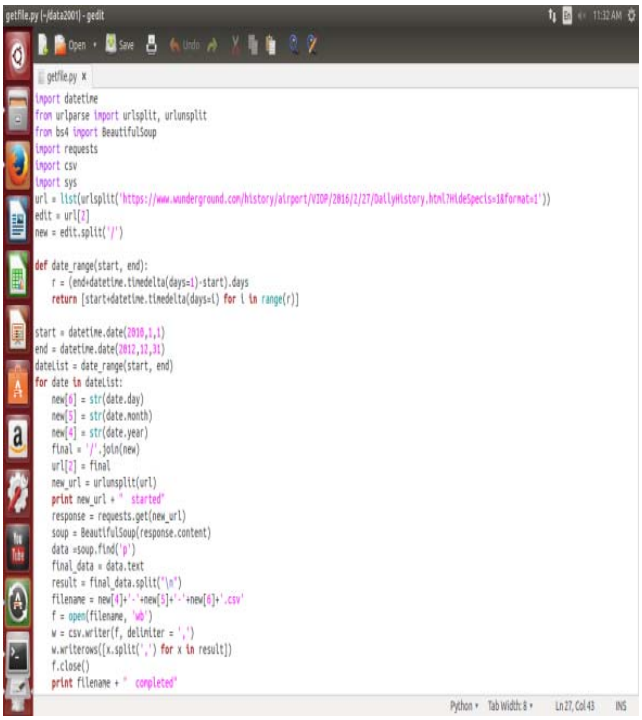


Fig. 1: Data Acquisition



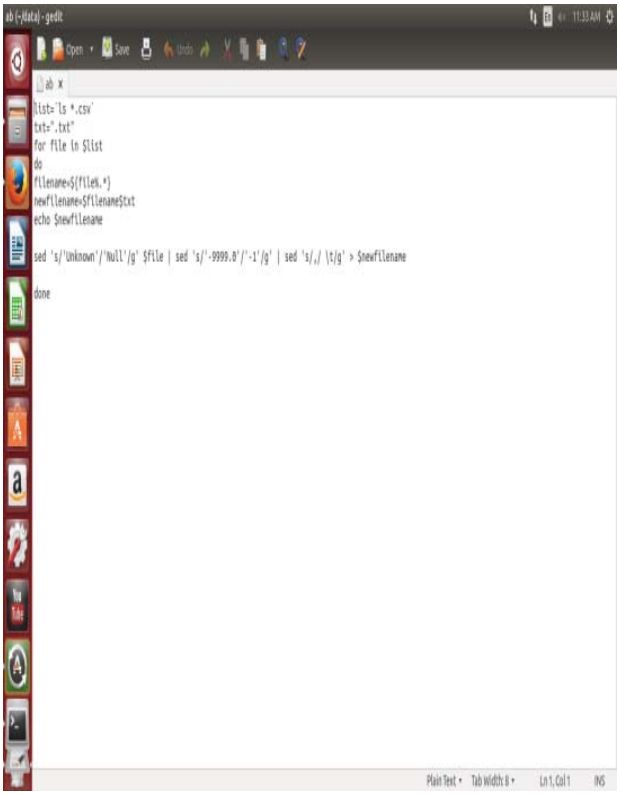
```
import datetime
from urlparse import urlsplit, urlunsplit
from bs4 import BeautifulSoup
import requests
import csv
import sys
url = list(urlsplit('https://www.wunderground.com/history/airport/VIDP/2016/2/27/DailyHistory.html?HideSpecis=1&format=1'))
edit = url[2]
new = edit.split('/')

def date_range(start, end):
    r = (end+datetime.timedelta(days=1)-start).days
    return [start+datetime.timedelta(days=i) for i in range(r)]

start = datetime.date(2010,1,1)
end = datetime.date(2012,12,31)
dateList = date_range(start, end)
for date in dateList:
    new[6] = str(date.day)
    new[5] = str(date.month)
    new[4] = str(date.year)
    final = '/'.join(new)
    url[2] = final
    new_url = urlunsplit(url)
    print new_url + "  started"
    response = requests.get(new_url)
    soup = BeautifulSoup(response.content)
    data =soup.find('p')
    final_data = data.text
    result = final_data.split("\n")
    filename = new[4]+'-'+new[5]+'-'+new[6]+'.csv'
    f = open(filename, 'wb')
    w = csv.writer(f, delimiter = ',')
    w.writerows([x.split(',') for x in result])
    f.close()
    print filename + "  completed"
```

Fig. 2: Screenshot of script for retrieving dataset

will convert it in txt on a single click. This data has some negative values and also missing values which have been replaced.



```
list= ls *.csv
txt=".txt"
for file in $list
do
filename=${file%.*}
newfilename=$filename$txt
echo $newfilename

sed 's/'Unknown'/'Null'/g' $file | sed 's/'-9999.0'/'-1'/g' | sed 's/,,/ \t/g' > $newfilename

done
```

Fig. 3: Screenshot of script for data conversion from CSV to TXT files

2.3 Data Set Description

The final dataset consists of five attributes for prediction of weather data. Out of five, four attributes are temperature, precipitation, humidity and sea level and fifth attribute is class attribute which shows the prediction of weather data into eight different classes. These classes are smoke, haze, fog, mist, rain, cloudy, clear and dust. The accuracy of weather dataset is evaluatedusing mean square error.

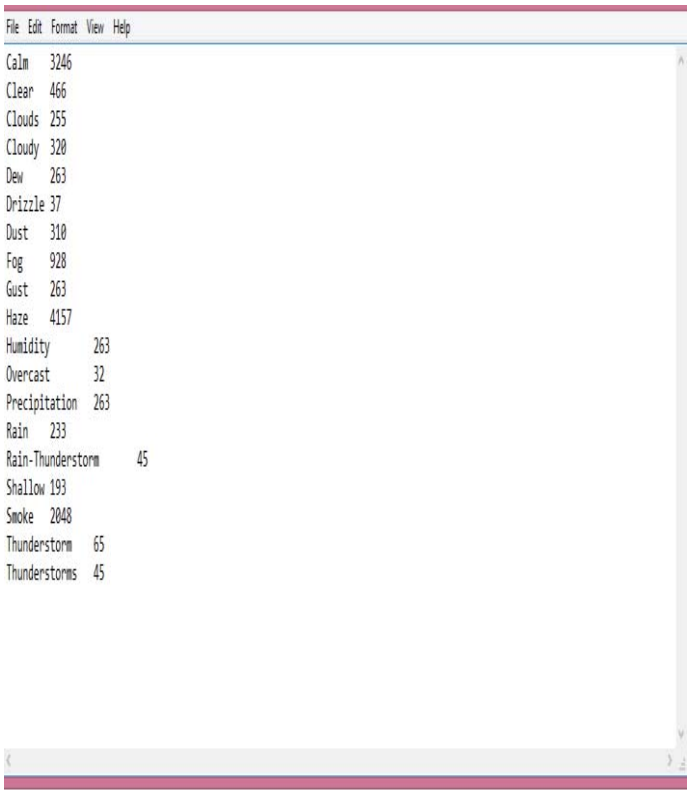
3. RESULTS AND DISCUSSION

3.1 Result of Wordcount in Hadoop

The wordcount algorithm reads text files and counts the number of times a word occurred. The input and output of wordcount algorithm is text files. The output of wordcountprogram is mentioned in Fig. 4.

3.2 Forecasting Using ANFIS and FL

ANFIS and FL are applied to forecasting of the weather data using MATLAB tool. The performances of these methods are compared on the basis of mean square error and the time taken parameters. The mean square error is 1.42087, 14.4194 and 16.8514 respectively. The time taken by fuzzy is longer than that of neural because of the rules generation process in fuzzy. Fig. 5 shows the output of fuzzy tool, whereas, figs. 6 shows the ANFIS model and 7 shows the results of ANFIS approach.



```
File  Edit  Format  View  Help
Calm    3246
Clear   466
Clouds  255
Cloudy  320
Dew     263
Drizzle 37
Dust    310
Fog     928
Gust    263
Haze    4157
Humidity        263
Overcast        32
Precipitation   263
Rain    233
Rain-Thunderstorm       45
Shallow 193
Smoke   2048
Thunderstorm    65
Thunderstorms   45
```

Fig. 4: Output of Wordcount

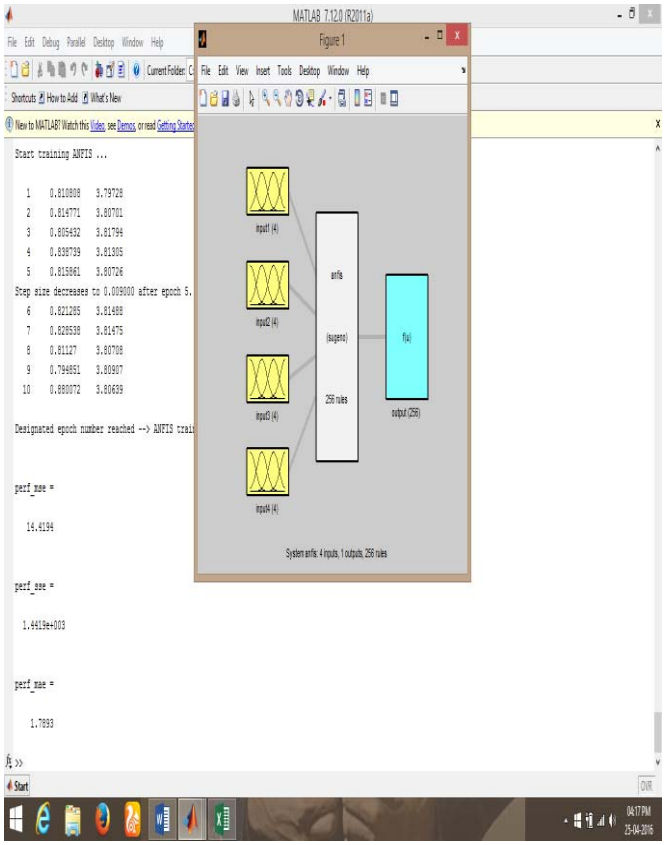


Fig. 5: Output of Fuzzy logic system

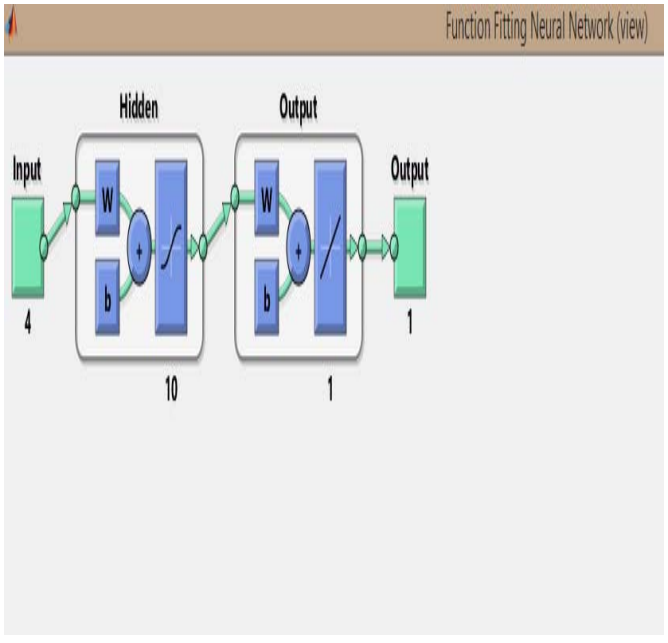


Fig 6:ANFIS structure

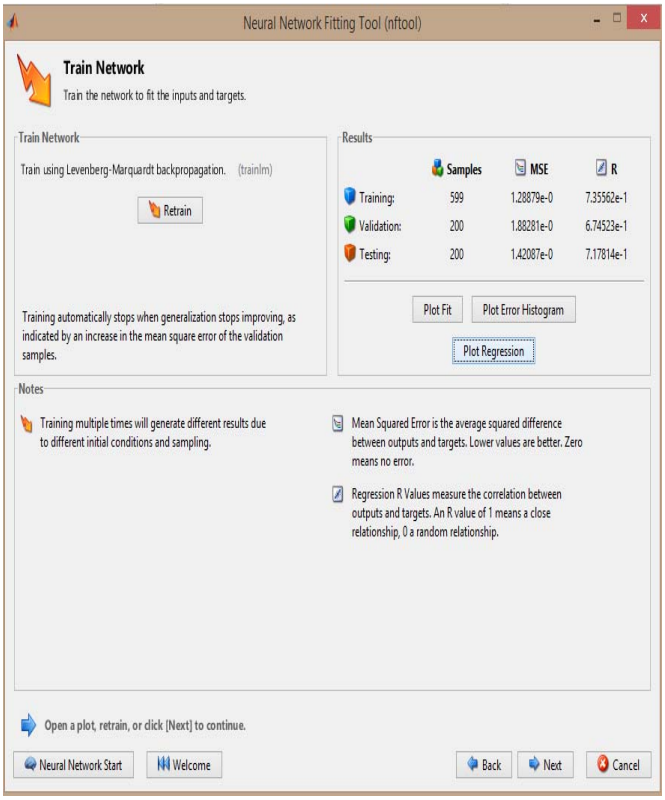


Fig 7: Output of ANFIS

Conclusion

In this work, an attempt is made to integrate the Hadoop tool with ANFIS and FL methods to forecast the weather data. Initially, weather data is collected from weather department website by using Beautiful SOUP and Python scripts. The collected data is pre-processed through Hadoop by using Wordcount algorithm. After the preprocessing, final dataset is obtained which is undergone for weather prediction process. For prediction of weather data, two data mining tool are applied i.e. ANFIS and FL. From results it is concluded that the ANFIS method predicts weather data more accurately. In future, other soft computing techniques are investigated for better prediction of weather data.

REFERENCES

[1] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, "Using Data-Mining for Short-Term Rainfall Forecasting", Notes in Computer Science, Volume 5518, 487-490

[2] Elia G. P., 2009, "A Decision Tree for Weather Prediction", Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1.

[3] Pielke R.A., "A comprehensive meterological modeling system RAMS," Meteorology and Atmosphe-ric Physics, Springer-Verlag Vol. 49, 69-91p,1992.

[4] Lutgens F.K., and Tarbuck E.J., The Atmospheric, 6th Edn.,Prentice Hall, Englewood Cliffs, NJ, 1995.

[5] Siddiqui Khalid J. and Nugen Steve M., Knowledge Based System for Weather Information Processing and Forecas-ting, Department of Computer Science, SUNY at Fredonia, NY14063, IEEE 1966.

[6] Sharma A., "A Weather Forecasting System using concept of Soft Computing: A new approach", PG Research Group SATI, Vidisha(M.P.), India, IEEE 2006.
[7] Khalid S., "Towards a Self-Configurable Weather Research and Forecasting System", School of Computing and Information Sciences, Florida International University, Miami FL, 2008.
[8] SenduruSrinivasulu, "Extracting Spatial Semantics in Association Rules for Weather Forecasting Image", Research Scholar Department of Information Technology, SathyabamaUniversity Chennai, India IEEE 2010.
[9] Wang Y. and Banavar S. "Convective Weather Forecast Accuracy Analysis at center and sector levels", NASA Ames Research center, Maffett Field, Califonia.
[10] Anad M. "Prediction and Classification of Thunderstorms using Artificial Neural Network", International Journal of Engineering Science and Technology (IJEST), Vol.3 (5) May 2011.