A Hadoop based Weather Prediction Model for Classification of Weather Data

Eesha S
B. Tech Computer Science, AI
Amrita Vishwa Vidypeeetham
Kollam, India
amenu4aie20025@am.students.amrita.edu

Gokul D. Raj
B. Tech Computer Science, AI
Amrita Vishwa Vidypeeetham
Kollam, India
amenu4aie20029@am.students.amrita.edu

Pavithra P M Nair
B. Tech Computer Science, AI
Amrita Vishwa Vidypeeetham
Kollam, India
amenu4aie20055@am.students.amrita.edu

R S Parvati Nair
B. Tech Computer Science, AI
Amrita Vishwa Vidypeeetham
Kollam, India
amenu4aie20058@am.students.amrita.edu

Raja Pavan Karthik
B. Tech Computer Science, AI
Amrita Vishwa Vidypeeetham
Kollam, India
amenu4aie20060@am.students.amrita.edu

***Abstract*—The use of technology to anticipate the weather at a specific location is known as weather forecasting. Big data has a significant impact on people's daily activities, such as research, and weather forecasts are also based on it. Data for weather forecasts comes from a variety of sources, including satellites and radar systems. The amount of data is enormous, and some of it is unstructured. It comprises both useful and worthless data for weather forecasting in the form of unstructured data. The overall state for that day is determined using the word count technique. Furthermore, methods for accurately forecasting meteorological data based on the mean square error include fuzzy logic (FL) and an artificial neural network with fuzzy interface system. The results of the experiments reveal that the ANFIS method produces more accurate findings than other methods.**

***Index Terms*—ANFIS, Fuzzy Logic, Weather, Hadoop**

I. Introduction

Weather forecasting is an important and vital activity in modern times, as it can effect numerous fields such as agriculture, animal husbandry, marine trade, and more, as well as saving the lives of sentient beings from climate risks, earthquakes, and other natural disasters. Temperature, precipitation, wind speed, wind direction, and humidity are all examples of atmospheric data that can be used to make a weather forecast, these parameters can be adjusted dynamically. Although there are a variety of technologies for forecasting weather data, the amount of data created for weather forecasting is huge and unstructured, as a result, predicting the weather using meteorological data is a difficult operation that necessitates a large number of criteria that can rapidly vary based on atmospheric circumstances. Various meteorological departments are cooperating to avert future weather-related climate concerns by exchanging information. Weather forecasting is a difficult task that may necessitate the use of cutting-edge technology and equipment in order to accurately predict the future. Predicting climatic conditions could be a difficult task for any living being to achieve. There is a need to study on meteorology to learn more about the weather. Meteorology is a multidisciplinary clinical study of the environment, including temperature, pressure, humidity, wind, and other variables. These variables can typically be measured with a thermometer, barometer, anemometer, and hygrometer. Multiple sensors positioned at a specific geographical location collect data on those attributes. This data is gathered by meteorological branches in a number of countries. This information is referred to as climate data.

II. Background Study

Big data analytics is the often complex process of analysing large amounts of data in order to identify perceptibly similar things like recurring patterns, correlations, request trends, and client preferences that can assist businesses in making well-informed decisions. Data analysis tools and methods allow businesses to examine data sets and obtain new information on a massive scale. Abecedarian questions concerning business operations and performance are answered by Business Intelligence (BI) queries. Powered by logical systems. Big data analytics will eventually lead to better and faster decision-making, modelling and prognostication of unborn issues, and improved corporate intelligence. Consider open-source software like Apache Hadoop, Apache Spark, and the entire Hadoop ecosystem as low-cost, flexible data storage and processing technologies built to handle the massive amounts of data created online when constructing your big data output. Big data is a collection of data from a range of sources that is typically described using the five qualities of volume, value, diversity, speed, and veracity.

The term ”Big Data” refers to a massive amount of information. The term ”volume” refers to a large amount of data. The term ”haste” relates to the rapid collection of information. Data comes in from a variety of sources in Big Data haste, including machines, networks, social media, and mobile phones. It refers to the three types of data: structured, semi-structured, and unstructured. Variety refers to the emergence of data from new sources both inside and outside an organisation. It relates to data inconsistencies and queries, i.e., data that is available might become untidy at times, and quality and delicacy are difficult to manage. After you’ve considered the four V’s, there’s one more V to consider: value. The majority of data with no value is useless to the organisation unless it is converted into a commodity that is useful.

Hadoop is an Apache open source framework written in Java that allows for the efficient processing of large datasets across clusters of computers using simple programming techniques. The Hadoop frame software functions as a girding that distributes garage and computation among clusters of computers. Hadoop is intended to scale from a single system to a large number of machines, each providing unique calculation and storage. The Hadoop armature is a combination of the train system, MapReduce, and HDFS . MapReduce/ MR1 or YARN/ MR2 are two types of MapReduce machines. Job Tracker, Task Tracker, NameNode, and DataNode are found in the master knot, while DataNode and TaskTracker are found in the slave knot.

III. Methodology and Dataset

A. ANFIS

An adaptive neuro-fuzzy conclusion system (ANFIS) is a sort of artificial neural network based on the Takagi- Sugeno fuzzy conclusion system. As it incorporates the concepts of neural networks and fuzzy sense, the trend was established in the early 1990s and has the latent ability to capture the benefits of both in a single frame. Its conclusion system is based on a collection of fuzzy IF-also rules that can compare nonlinear functions with literacy. As a result, ANFIS is regarded as a universal estimator. The stylish parameters obtained by inheritable algorithms can be used to employ ANFIS more efficiently and optimally. It’s a component of the intelligent situation-apprehensive energy management system.

B. Fuzzy Logic

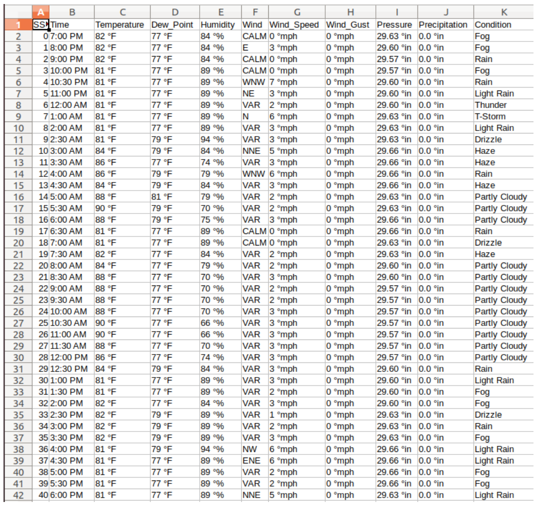
Fuzzy Sense is a type of multivalued sense and the verity value of variables can be any real number between zero and one. It’s used to deal with the concept of partial verity, in which the verity value can vary from true to false. The verity values of variables in discrepancy can only be the integer numbers 0 or 1 in a Boolean meaning. Scientist Lotfi Zadeh proposed the fuzzy set thesis in 1965, which gave rise to the term fuzzy sense. Still, fuzzy sense has been investigated as a horizonless sense of values since the 1920s, particularly by Ukasiewicz and Tarski. The concept of fuzzy sense is based on the observation that humans develop views based on squishy, non-numerical data. Models, sometimes known as fuzzy sets, are a good way to describe ambiguous and squishy data. These models can be used to celebrate, represent, manipulate, analyse, and utilise data and information that is imprecise or unclear. From control proposition to artificial intelligence, fuzzy sense has been used in a variety of disciplines.

A Hadoop-based weather forecasting model is proposed in this project for efficient processing and forecasting of meteorological data. Following the steps below, the entire procedure was completed:

1) First, data is collected from the source Wunderground in order to predict weather data.
2) The gathered data is preprocessed using the HADOOP word count technique.
3) The dataset’s categorization attributes have been identified and documented.
4) ANFIS and fuzzy logic were used for forecasting.
5) Analyze the outcomes.

C. Data Acquisition

The HTML of this page is processed in order to collect the desired data, and the data is stored in.csv format. Beautiful Soup will parse the HTML code and retrieve the data. It’s a Python package for parsing HTML and XML files and extracting data from them. It collaborates with a parser to offer idiomatic navigation, searching, and modification of the parse tree. Meteorological data spanning ten years is gathered in this study. A script using Python and BeautifulSoup is created to obtain all of the data. The date and time are used to create a date record. Date and Time is a simple to import and use library. The URL for a specific tag is generated using urlparse, urlsplit, and urlunsplit, and then parsed with BeautifulSoup. The output is saved as a CSV (Comma Separated Values) file.



	A	B	C	D	E	F	G	H	I	J	K
1	SS	Time	Temperature	Dew_Point	Humidity	Wind	Wind_Speed	Wind_Gust	Pressure	Precipitation	Condition
2	0	7:00 PM	82 °F	77 °F	84 °%	CALM	0 °mph	0 °mph	29.63 °in	0.0 °in	Fog
3	1	8:00 PM	82 °F	77 °F	84 °%	E	3 °mph	0 °mph	29.60 °in	0.0 °in	Fog
4	2	9:00 PM	82 °F	77 °F	84 °%	CALM	0 °mph	0 °mph	29.57 °in	0.0 °in	Rain
5	3	10:00 PM	81 °F	77 °F	89 °%	CALM	0 °mph	0 °mph	29.57 °in	0.0 °in	Fog
6	4	10:30 PM	81 °F	77 °F	89 °%	WNW	7 °mph	0 °mph	29.60 °in	0.0 °in	Rain
7	5	11:00 PM	81 °F	77 °F	89 °%	NE	3 °mph	0 °mph	29.60 °in	0.0 °in	Light Rain
8	6	12:00 AM	81 °F	77 °F	89 °%	VAR	2 °mph	0 °mph	29.60 °in	0.0 °in	Thunder
9	7	1:00 AM	81 °F	77 °F	89 °%	N	6 °mph	0 °mph	29.63 °in	0.0 °in	T-Storm
10	8	2:00 AM	81 °F	77 °F	89 °%	VAR	3 °mph	0 °mph	29.63 °in	0.0 °in	Light Rain
11	9	2:30 AM	81 °F	79 °F	94 °%	VAR	3 °mph	0 °mph	29.63 °in	0.0 °in	Drizzle
12	10	3:00 AM	84 °F	79 °F	84 °%	NNE	5 °mph	0 °mph	29.66 °in	0.0 °in	Haze
13	11	3:30 AM	86 °F	77 °F	74 °%	VAR	3 °mph	0 °mph	29.66 °in	0.0 °in	Haze
14	12	4:00 AM	86 °F	79 °F	79 °%	WNW	6 °mph	0 °mph	29.66 °in	0.0 °in	Rain
15	13	4:30 AM	84 °F	79 °F	84 °%	VAR	3 °mph	0 °mph	29.66 °in	0.0 °in	Haze
16	14	5:00 AM	88 °F	81 °F	79 °%	VAR	2 °mph	0 °mph	29.63 °in	0.0 °in	Partly Cloudy
17	15	5:30 AM	90 °F	79 °F	70 °%	VAR	2 °mph	0 °mph	29.63 °in	0.0 °in	Partly Cloudy
18	16	6:00 AM	88 °F	79 °F	75 °%	VAR	3 °mph	0 °mph	29.66 °in	0.0 °in	Partly Cloudy
19	17	6:30 AM	81 °F	77 °F	89 °%	CALM	0 °mph	0 °mph	29.66 °in	0.0 °in	Rain
20	18	7:00 AM	81 °F	77 °F	89 °%	CALM	0 °mph	0 °mph	29.63 °in	0.0 °in	Drizzle
21	19	7:30 AM	82 °F	77 °F	84 °%	VAR	2 °mph	0 °mph	29.63 °in	0.0 °in	Haze
22	20	8:00 AM	84 °F	77 °F	79 °%	VAR	2 °mph	0 °mph	29.60 °in	0.0 °in	Partly Cloudy
23	21	8:30 AM	88 °F	77 °F	70 °%	VAR	3 °mph	0 °mph	29.60 °in	0.0 °in	Partly Cloudy
24	22	9:00 AM	88 °F	77 °F	70 °%	VAR	2 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
25	23	9:30 AM	88 °F	77 °F	70 °%	VAR	2 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
26	24	10:00 AM	88 °F	77 °F	70 °%	VAR	3 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
27	25	10:30 AM	90 °F	77 °F	66 °%	VAR	3 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
28	26	11:00 AM	90 °F	77 °F	66 °%	VAR	3 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
29	27	11:30 AM	88 °F	77 °F	70 °%	VAR	3 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
30	28	12:00 PM	86 °F	77 °F	74 °%	VAR	3 °mph	0 °mph	29.57 °in	0.0 °in	Partly Cloudy
31	29	12:30 PM	84 °F	79 °F	84 °%	VAR	3 °mph	0 °mph	29.60 °in	0.0 °in	Rain
32	30	1:00 PM	81 °F	77 °F	89 °%	VAR	3 °mph	0 °mph	29.60 °in	0.0 °in	Light Rain
33	31	1:30 PM	81 °F	77 °F	89 °%	VAR	2 °mph	0 °mph	29.60 °in	0.0 °in	Fog
34	32	2:00 PM	82 °F	77 °F	84 °%	VAR	3 °mph	0 °mph	29.60 °in	0.0 °in	Fog
35	33	2:30 PM	82 °F	79 °F	89 °%	VAR	1 °mph	0 °mph	29.63 °in	0.0 °in	Drizzle
36	34	3:00 PM	82 °F	79 °F	89 °%	VAR	2 °mph	0 °mph	29.63 °in	0.0 °in	Rain
37	35	3:30 PM	82 °F	79 °F	89 °%	VAR	3 °mph	0 °mph	29.63 °in	0.0 °in	Fog
38	36	4:00 PM	81 °F	79 °F	94 °%	NW	6 °mph	0 °mph	29.66 °in	0.0 °in	Light Rain
39	37	4:30 PM	81 °F	77 °F	89 °%	ENE	6 °mph	0 °mph	29.66 °in	0.0 °in	Light Rain
40	38	5:00 PM	81 °F	77 °F	89 °%	VAR	2 °mph	0 °mph	29.66 °in	0.0 °in	Fog
41	39	5:30 PM	81 °F	77 °F	89 °%	VAR	2 °mph	0 °mph	29.66 °in	0.0 °in	Fog
42	40	6:00 PM	81 °F	77 °F	89 °%	NNE	5 °mph	0 °mph	29.63 °in	0.0 °in	Light Rain

Fig. 1. Dataset retrieved after Webscraping

D. Data Pre-processing

The word count algorithm in TXT format is used in HADOOP for data processing. A bash script is used to convert the data files for each day to TXT format. The script can be used to convert a whole folder of data to TXT with a single click. There are some negative values in this data, as well as missing numbers that have been rectified.

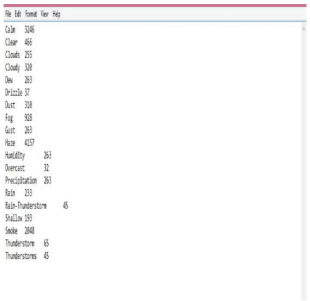
E. Dataset Description

The resulting data set has five attributes for weather forecasting. Temperature, precipitation, humidity, and sea level are the first four attributes, and the fifth attribute is a class attribute that indicates expected weather data. Smoke, Fog, Haze, Fog, Rain, Cloudy, Clear, and Dust are the different grades of this class. The mean square error is used to determine the correctness of the meteorological data collection..

IV. Experiment Results and Analysis

A. Result Of WordCount in Hadoop

The word counting algorithm examines text files and counts how many times a word appears. Here, the role of Mapper is to map the keys to the existing values and the role of Reducer is to aggregate the keys of common values. So, everything is represented in the form of Key-value pair. The word count algorithm takes text files as input and output.

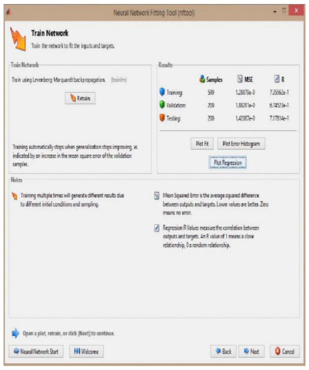


```
Calm    3246
Clear   466
Clouds  255
Cloudy  320
Dew     263
Drizzle 37
Dust    310
Fog     928
Gust    263
Haze    4157
Humidity        263
Overcast        32
Precipitation   263
Rain    233
Rain-Thunderstorm       45
Shallow 193
Smoke   2048
Thunderstorm    65
Thunderstorms   45
```

Fig. 2. Output of WordCount

B. Forecasting using ANFIS and FL

Using the MATLAB programme, ANFIS and FL are used to forecast meteorological data. ANFIS is an integration system in which neural networks are applied to optimize the fuzzy inference system. ANFIS constructs a series of fuzzy if–then rules with appropriate membership functions to produce the stipulated input–output pairs. Fuzzy logic is an approach to computing based on ”degrees of truth” rather than the usual ”true or false” Boolean logic on which the modern computer is based. It’s used to deal with the concept of partial verity, in which the verity value can vary from true to false.In artificial intelligence (AI) systems, fuzzy logic is used to imitate human reasoning and cognition. Rather than strictly binary cases of truth, fuzzy logic includes 0 and 1 as extreme cases of truth but with various intermediate degrees of truth. The verity values of variables in discrepancy can only be the integer numbers 0 or 1 in a Boolean meaning. The mean square error and the time parameters utilised are used to compare the performance of these approaches. The mean square error is 14.4194 and 16.8514 correspondingly. Due to the rule generation process of Fuzzy, it takes longer than Neural.



Train Network

Train the network to fit the inputs and targets.

Train Network

Train using Levenberg-Marquardt backpropagation. (trainlm)

Training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples.

Results

	Samples	MSE	R
Training:	599	1.28879e-0	7.35562e-1
Validation:	200	1.88281e-0	6.74523e-1
Testing:	200	1.42087e-0	7.17814e-1

Notes

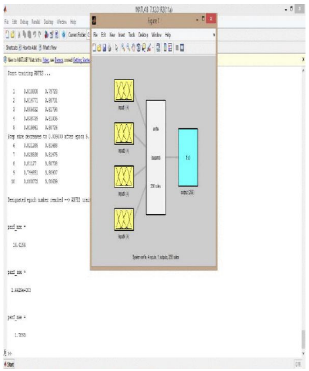
Training multiple times will generate different results due to different initial conditions and sampling.

Mean Squared Error is the average squared difference between outputs and targets. Lower values are better. Zero means no error.

Regression R Values measure the correlation between outputs and targets. An R value of 1 means a close relationship, 0 a random relationship.

Open a plot, retrain, or click [Next] to continue.

Fig. 3. Output of ANFIS


MATLAB 7.12.0 (R2011a)
File Edit Debug Parallel Desktop Window Help
Current Folder
Shortcuts How to Add What's New
Figure 1
File Edit View Insert Tools Desktop Window Help
input1 (4)
input2 (4)
input3 (4)
input4 (4)
anfis
(sugeno)
256 rules
f(u)
output (256)
System anfis: 4 inputs, 1 outputs, 256 rules
Start

```
Start training ANFIS ...

   1    3.810808    3.79728
   2    3.814771    3.80701
   3    3.815832    3.81796
   4    3.818735    3.81836
   5    3.819862    3.80726
Step size decreases to 0.009033 after epoch 5.
   6    3.821285    3.81688
   7    3.828538    3.81475
   8    3.81127     3.80708
   9    3.794851    3.80807
  10    3.808072    3.80639

Designated epoch number reached --> ANFIS trai


perf_mse =

   16.8514


perf_mae =

   1.4415e+003


perf_sae =

   1.7893

>>
```

Fig. 4. Output of fuzzy logic system

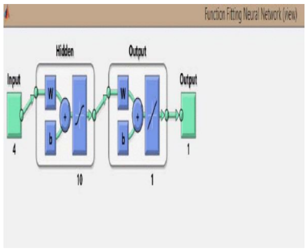


Fig. 5. Structure of ANFIS

V. Conclusion and Future Work

To forecast meteorological data, an effort is made to connect the Hadoop tool with ANFIS and FL approaches in this paper. Weather data is initially acquired through Beautiful SOUP and Python scripts from the weather department’s website. The obtained data is pre-processed using the Wordcount technique in Hadoop. After preprocessing, the finalized dataset is created, which is then used in the weather forecasting process. Two data mining tools, ANFIS and FL, are used to forecast meteorological data. The ANFIS technique forecasts meteorological data more correctly, according to the findings. Other soft computing techniques will be researched in the future for better weather data prediction.

Acknowledgment

We would like to express our sincere gratitude to Sanjay S Kumar, for guiding us to do our work and report. We would also like to express our sincere gratitude to our beloved college Amrita Vishwa Vidypeetham for providing and presenting us an opportunity to work on this report.

References

[1] A. K. Pandey, C. P. Agrawal and M. Agrawal, ”A hadoop based weather prediction model for classification of weather data,” 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017, pp. 1-5, doi: 10.1109/ICECCT.2017.8117862.

[2] Priyanka Chouksey, and Abhishek Singh Chauhan. (n.d.). A Review of Weather Data Analytics using Big Data. International Journal of Advanced Research in Computer and Communication Engineering. Retrieved June 3, 2022, from https://ijarcce.com/

[3] Ning Yang, Lewis Westfall, Ms. Preeti Dalvi. (n.d.). A Weather Prediction Model with Big Data. PACE UNIVERSITY. Retrieved June 3, 2022, from https://www.pace.edu/seidenberg

[4] Adaptive neuro fuzzy inference system - Wikipedia. (n.d.). Adaptive Neuro Fuzzy Inference System - Wikipedia; en.wikipedia.org. Retrieved June 3, 2022, from https://en.wikipedia.org/wiki/Adaptive neuro fuzzy inference system

[5] Apache Hadoop. (n.d.). Apache Hadoop; hadoop.apache.org. Retrieved June 3, 2022, from https://hadoop.apache.org/

[6] Big Data Analytics IBM.(n.d.).Big Data Analytics IBM; www.ibm.com. Retrieved June 3, 2022, from https://www.ibm.com/analytics/big-data-analytics

[7] Big data - Wikipedia. (2016, April 13). Big Data - Wikipedia; en.wikipedia.org. https://en.wikipedia.org/wiki/Big data