WEATHER PREDICTION: IMPROVING ACCURACY USING DATA MINING AND FORECASTING TECHNIQUES

PEDRO ALEJANDRO MARQUEZ

Master's Program in Industrial Engineering

APPROVED:

Jose F. Espiritu, Ph.D., Chair

Heidi A. Taboada, Ph.D.

Virgilio Gonzalez, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

Dedication

I dedicate every effort I made to complete this thesis project to my family. A special feeling of gratitude to my parents, Pedro and Blanca whose affection, love, encouragement and prayers day and night made me able to achieve success and honor them. Thank you to my brothers, Carlos and Andres, whose words of encouragement made it possible. Thank you to the staff and faculty that have guided me along my graduate journey.

WEATHER PREDICTION: IMPROVING ACCURACY USING DATA MINING AND FORECASTING TECHNIQUES

by

PEDRO ALEJANDRO MARQUEZ, B.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Industrial, Manufacturing and Systems Engineering

THE UNIVERSITY OF TEXAS AT EL PASO

August 2020

Acknowledgements

I want to thank first and foremost my advisor Dr. Jose Espiritu for all of his support, patience, guidance, and understanding. Thank you Dr. Taboada for guiding me along the way and helping me achieve my goals. Also, I would also thank Pablo Bustamante for his guidance and sharing of his knowledge on data mining during the project.

Special thanks to Dr. Irma Lawrence, who is responsible for the USDA's NIFA HSI program, for giving me the opportunity to carry out this higher education program. Finally, I thank the University of Texas at El Paso for providing me with the tools I needed to achieve academic success.

Abstract

This study is focused to provide the insights of weather to understand the significance of weather changes in any parameter. Weather forecasting contributes to the social and economic welfare in many sections of the society. Weather is extremely difficult to predict because it is a complex and chaotic system. This means that small errors in the initial conditions of a forecast grow rapidly and affect predictability. Nowadays, massive real-time data is being generated by IoT devices, radars, weather stations, and satellites. The need to adopt big data analytics in IoT applications is compelling. These two technologies have already been recognized in the fields of IT and business. Data mining techniques and machine learning algorithms need to be considered and trained with big data to improve the accuracy of weather forecasts. The contribution to this problem is to analyze the accuracy and correlation between weather conditions with the use of different data mining and forecasting techniques to predict precipitation for the next year.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Chapter 1 will investigate the integration of two emerging concepts in the fields of information technology and business: 'Internet of Things' and 'Big Data'. Each concept was examined by their impact and their development of smart connected communities. This chapter analyses how the world has been revolutionized by these technologies by identifying and analyzing several opportunities and challenges presented by the capabilities to ingest and utilize huge amounts of 'Internet of Things' data, which includes applications in smart cities, smart agriculture, smart transportation, etc. Finally, this chapter approaches how the generation of real-time data from the 'Internet of Things' devices created an opportunity to apply data mining techniques and develop smart weather forecasts.

1.1 Introduction

In future years, it is expected that cities will face several challenges such as safety, sustainability, energy use, effective service delivery, effective transportation systems, etc. Advances in the network integration of information systems, sensing and communication devices, data sources, and artificial intelligence, are generating opportunities to develop smart and connected communities. Nowadays, people expect the power of sensors, cloud computing, high speed networks and data analytics with the use of a myriad of things like smart phones, cars, social networks, and transportation apps like Uber.

According to Cheng et al. (2016), statistics show that 500 billion devices are expected to be connected to the Internet by 2030. Each device includes sensors that collect data, interact with the environment, and communicate over the network without any manual intervention with the help of embedded technology. The Internet of Things (IoT) is the network of these connected devices.

On September 14, 2015, the United States government announced a new Smart Cities Initiative to support local communities tackle key challenges such as reducing traffic congestion, fighting crime, fostering economic growth, managing the effects of a changing climate, and improving the delivery of city services. On November 25, 2015, the Networking and Information Technology Research and Development (NITRD) Program announced the release of version 4 of a Smart and Connected Communities Framework (NITRD, 2015). The framework outline is that communities in all settings and at all scales have access to Internet of Things technologies and services to improve the health, safety and economy of their residents.

IoT is the primary enabler of a larger industry transformation called digital business. Digital business is one that uses technology as an advantage in its internal and external operations. Sensors can detect location, environment, presence, and more. Big Data Analytics (BDA) is emerging as a key to analyzing IoT generated data from "connected devices" which helps to take the initiative to improve decision making and provide competitive advantage. IoT securely connects devices and fuels applications that can be delivered as services. Organizations are increasingly looking to digital technologies to create or enhance their business models, processes and services; to empower workforce efficiency and innovation; and personalize the citizen, customer, or employee experience.

1.2 INTERNET OF THINGS

Today, people are almost completely dependent on computers and the internet for information. The problem arises when human beings have limited time, attention, and accuracy at capturing data about our surroundings. As a result, Internet of Things is providing a linked set of computer programs and sensors that do not experience the same limitations than people do. Internet

of Things (IoT) is an interconnection of several "smart devices" that are equipped with microchips, sensors and wireless communication capabilities to achieve a common goal.

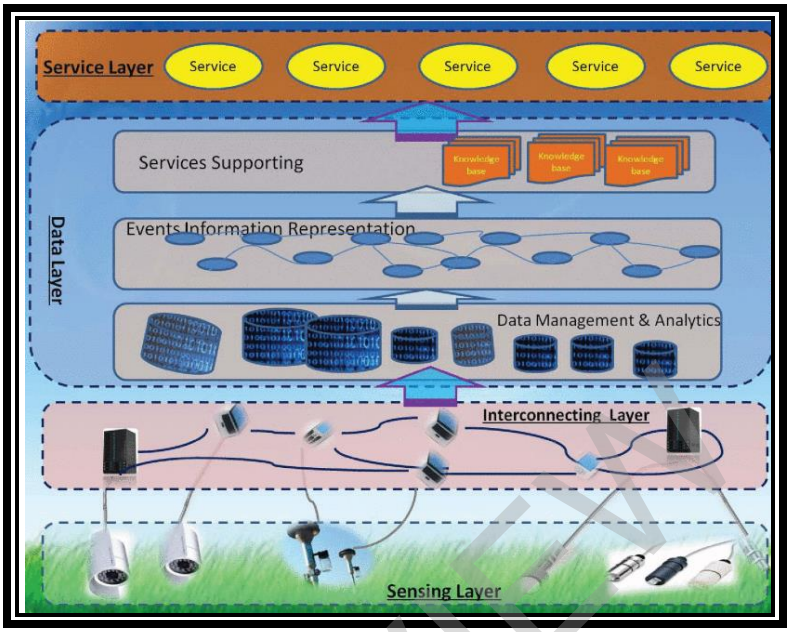

Figure 1.1: Function Architecture of Internet of Things for Smart (Sun et al. 2016).

The architecture of IoT consists of four different layers: sensing layer, interconnecting layer, data layer, and service layer (Sun et al. 2015). The sensing layer's purpose is to perform ubiquitous sensing. Some examples of the most utilized sensors are smart phones, cameras, audios, accelerometers, GPS, gyroscope, compass, proximity, and ambient light. The interconnecting layer has the purpose of transmitting data and exchanging information among devices. Any IoT device is assigned an IP address, which identifies the network that the device is in. The Data layer is responsible for structuring and analyzing massive, trivial, heterogeneous data generated from different kinds of monitoring devices in the sensing layer. Meaningful and useful information is extracted from the data and represented in an efficient way that provides service support and actionable intelligence for the user. Finally, the service layer, or application layer, provides various services based on the data analyses.

As the number of connected IoT devices continues to grow, the amount of data generated by these devices will also grow. Some devices might just produce a minimum amount of data indicating only one metric, while others like video surveillance cameras generate huge amounts of data to examine, such as crowds of people surveillance. Immense data sets (petabytes or gigabytes) can demand advanced forms of information processing for better insights and decision making.

1.3 BIG DATA

Big Data has become important for any organization that generates a huge amount of heterogeneous data, which if captured, processed, and analyzed will reveal patterns and provide insights. Big Data concept emerged when large volumes of structured, semi-structured, and unstructured data posed a difficult task for processing using traditional methods and databases.

The size, speed, and format in which data is generated affect the quality of the information. Data can come from different sources such as business transaction systems, customer databases, mobile applications, websites, machines and real-time data sensors produced in IoT systems. This comes with challenges defined as 5Vs i.e. Volume, Variety, Velocity, Veracity, Value (Juneja & Das, 2019).

Gantz & Reinsel (2012) characterizes big data into three aspects: data sources, data analytics, and the presentation of the results of the analytics. Volume indicates the size of the data sets created at high frequency rates. Variety is present when there are different types of data types, such as structured, semi-structured or unstructured data. Velocity refers to the speed and frequency at which the data is created. Veracity deals with the accuracy, truthfulness and authenticity of the data. Value denotes the worthiness of data extracted from various raw data available, in other words, not all data extracted is useful.

According to Juneja & Das (2019), data flows through four phases in the Big Data System Lifecycle. These phases are Data Origin Identification, Data Acquisition, and Cleansing, Data Aggregation and Storage and Data Analysis as presented in Figure 1.2.

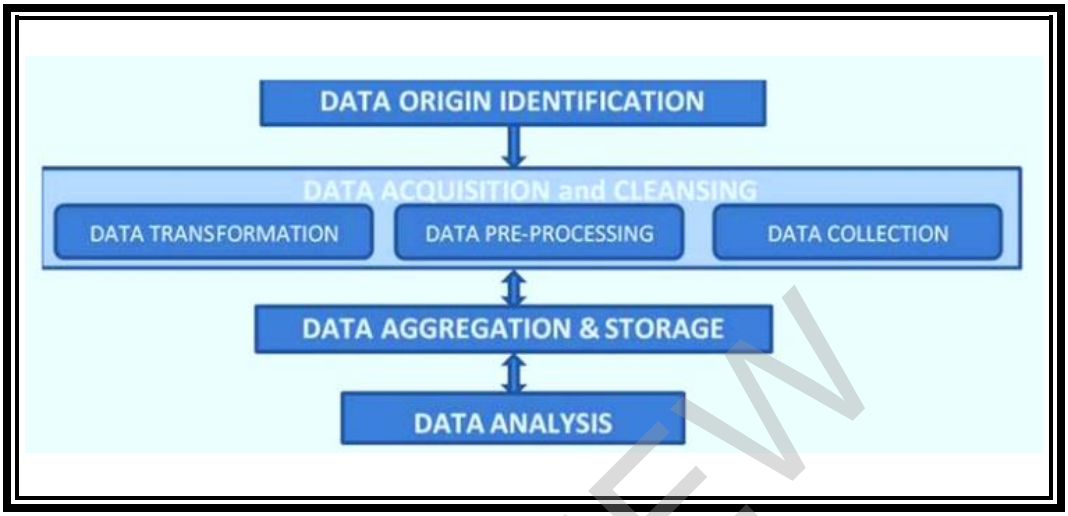

Figure 1.2: Big Data Lifecycle (Juneja & Das, 2019).

The first phase is ***Data Origin Identification***. This phase reviews the raw data being produced from multiple sources. The sources may include social sites, financial applications, customer relation applications, media web sites, images, etc. It is crucial to understand the veracity and reliability of data sources.

Secondly, ***Data Acquisition and Cleansing*** phase comprehends the data from many sources. This phase deals with the variety and value complexities. The raw data can be collected with anomalies, badly formatted, or a combination of structured, semi-structured and unstructured data. Such data requires to be cleaned and filtered, reformatted and structured, remove illegal values and compressed. These pre-processing steps are crucial to transform the data and eliminate the 5Vs variety and value complexities for analysis.

The third phase is ***Data Aggregation and Storage*** which confirms that all heterogeneous sources are stored and joined across databases.

Finally, the ***Data Analysis*** phase compares data characteristics to identify patterns usually by using high-level programming skills and methodologies. The results of this phase should inform the user about the final state, make forecasts, or provide an approach for decision making.

1.4 BIG DATA ANALYTICS IN IOT

The exponential growth of data produced by IoT devices has played a major role in the Big Data field. Big data analytics is quickly emerging as a significant IoT initiative to improve decision making. Big data analytics refers to the process of collecting, storing and analyzing large volumes of data sets to reveal trends, unseen patterns, hidden correlations, and new information (Golchha 2015). Big data analytics in IoT demands storing the data in several storage technologies. Big data implementations will require performing lightning-fast analytics with queries to allow organizations to gain rapid insights and make quick decisions. Depending on the requirements of the developed IoT applications, different analytic types have been discussed in this subsection under real-time, offline, business intelligence level, and massive level analytics (Chen & Zhang 2014). Furthermore, a comparison based on analytics types and their level is shown in Table 1.1.

Table 1.1: Comparison of Different Analytics Types and Their Levels (Marjani et al., 2017).

Analytic Type	Specified Use	Advantages
Real Time	To analyze the large amounts of data generated by the sensors	Parallel processing clusters using traditional databases memory-based computing platforms
Offline	To use for the Applications where there is no high requirements on response time.	Efficient data acquisition Reduce the cost of data format conversion
Memory Level	To use where the total data volume is smaller than the maximum. Memory of the cluster	-Real Time
Business Intelligence Data	To use when data scale surpasses the memory level	Both offline and online
Massive level	To use when data scale totally surpasses the capacity of business intelligence products and traditional databases.	Most are Offline

Historical data analysis uses a set of historical data for batch analysis. Real-time analytics instead visualizes and analyzes the data as it appears in the computer system. The role of big data

analytics in IoT is to process a massive volume of data on a real-time basis and take out the value by enabling high-velocity capture, discovery, and analysis. IoT big data processing follows four sequential steps: collecting, storing, mining and visualization. These four steps are described below.

1. Data generated by IoT devices is collected in the big data system. Big data is a large quantity of data characterized by the 3V's (volume, variety, velocity) definition proposed by Bayer (2011). Volume is the quantity of data produced. Variety refers to various forms of information that are retrieved like voice, texts, images, videos, document, sensor data, tweets, etc. Lastly, velocity approaches the high-speed at which these data is generated.
2. In the big data system, which is mainly a shared distributed database, a tremendous amount of data is stored in big data files. Big Data requires a parallel and distributed system architecture to store this data. Since data is provided from different sites and in different places, it needs to be stored in uniform manner.
3. Internet of Things data stored is used for analytics. The data from various sources is collected; they are refined and stored under uniform schemas. IoT applications involve data sets that may have a varied structure as unstructured, semi-structured and structured data sets. There may also be a significant difference in the data formats and types. Analyzing these large data allows people to discover the correlation, facts and other important information that lies in this large data set, which is impossible to be determined by human.
4. Finally, big data analytics generates reports, tables, graphs, diagrams or applications that uncover hidden patterns, unknown correlations, market trends, customer preferences or other useful business information. Usually, it allows the business executive to analyze all these varying sets of data using automated tools and software.

1.5 Applications, Features, and Products of IoT Data Analytics in Agriculture

A variety of IoT-based applications are being used in different sectors such as smart cities, smart home, smart business, agriculture, transportation, healthcare, logistics, etc. These applications have succeeded in providing enormous benefits to the users. These days, these applications have steadily increased and some of them are already deployed and being used at different levels (Want et al, 2015). IoT's applications require hardware, middleware and presentation. These applications have features such as interaction with the environment, interaction between people and devices, automatic routine tasks with less supervision, self-organized infrastructure and communication security (Javed et al 2018). Smart agriculture has been one of the most developed fields from IoT.

According to Elijah (2018), the world population is estimated to be about 9.7 billion in 2050 as such there will be great demand for food. Data analytics and IoT devices enables better and smarter agriculture that will allow people to overcome this demand. Some of the opportunities for implementing IoT devices in the agricultural sector are crop and livestock, machinery, irrigation and water quality monitoring, weather monitoring, soil monitoring, disease and pest control, automation and precision. Developing IoT and data analytics applications in the agricultural field enhance farmers' productivity, quality and profit.

The agriculture supply chain is composed of six different activities as shown in Figure 1.3. These six activities are input suppliers, farms, traders, processors, retailers, and consumers.

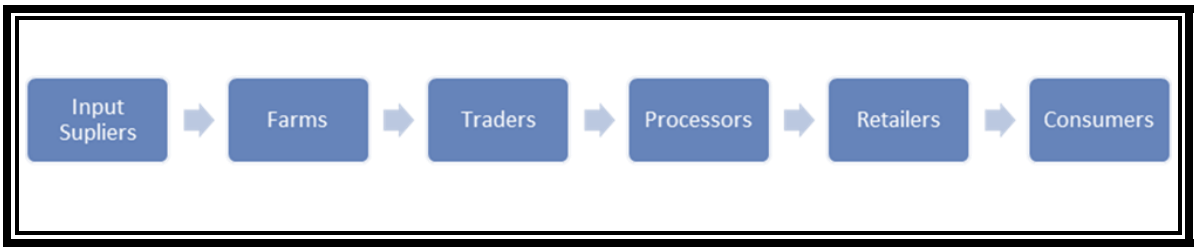

Figure 1.3: Supply Chain in Agriculture

IoT provides opportunities throughout the different activities in the supply chain in order to gain competitive advantage. For example, smart farms can offer insight on resource management. With the implementation of IoT and data analytics, farmers can make more effective purchases by optimizing and forecasting the use of machinery, fertilizers, seeds and chemicals from suppliers.

Sensors applied in smart agriculture generate data that assist farmers to monitor and optimize crops by adapting to changes in the environmental conditions. These sensors are placed on weather stations, drones, and machinery in the agriculture industry. Sensors provide valuable information on moisture soil, humidity levels, trunk diameter of plants, microclimate condition, as well as to forecast weather. Automatic climate control according to harvesting requirements, timely and controlled irrigation, and humidity control for fungus prevention are examples of actions performed based on big data analytics recommendations.

Agricultural gathering vehicles have been equipped with sensors, wireless communication, and dynamic programming has allowed the creation of autonomous vehicles. Furthermore, IoT sensors has been installed in the vehicles generate information that can be used to track current location and analyze efficiency in delivery times, fuel consumption, or delivery routes. IoT devices installed within the agricultural equipment will provide constant, accurate measurements of output to isolate sources of waste, gain process control, maximize productivity and ensure quality.

Connected devices and products gives retail companies the opportunity to optimize operations in their supply chain and improve the customer experience. One example of IoT technologies is RFIDs, which can precisely track inventory. Data visualization technologies allow employees to track products across the supply chain. Retail companies also count on the Internet of Things application development to improve self-checkout. Identifying person monitoring traffic

patterns in stores and trying to find a connection with trends can provide accurate data about how customers behave.

Like companies, government agencies are trying to deliver quality services in multifaceted environments. Smart infrastructure technologies can allow government agencies planners to measure and monitor traffic management, security, waste, energy, and water supplies to lower costs and improve services for the citizens (Meyers et al., 2015). At a federal level, agencies are more focused on scaling measurement capabilities: The Department of Defense uses RFID chips to monitor its supply chain (Defense Industry Daily, 2010), the US Geological Survey uses IoT devices to monitor the bacterial levels of rivers and lakes (Meyers et al., 2015), and the General Services Administration has initiated employing sensors to measure the energy efficiency of "green" buildings (Fowler et al., 2015). Regarding changes in the weather, IoT devices will provide information about temperatures, noise and air pollution to several agencies such as the National Weather Service and the Department of Agriculture.

Weather warning and forecasting applications seems to be one of the most useful applications for the government, industry, and the general public. The United States National Weather Service (NWS) supports all aspects of keeping the public safe from weather, water, and climate hazards by providing weather warnings and forecasting programs (weather.gov). In the agricultural field, sensors are retrieving useful information to forecast weather. Data analytics helps farmers to reduce the probability of having production risks and to create mitigation tactics for unexpected events. In addition, data analytics and IoT devices can significantly influence decisions and policymaking for food security.

1.6 Significance of Weather Forecasting

Weather forecasting is one of the most prominent topics that has been challenging scientists and engineers due to an expected increase of weather events and climate changes around the world. Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Human beings have attempted to predict the weather since ancient times and started developing scientific forecast models since the nineteenth century. Nowadays, forecasters use more advanced methods and technologies to gather weather data, along with the world's most powerful computers. With the ability to launch satellites and supercomputers and harvest data from IoT devices, the new arrivals are advancing in information-gathering capabilities. Also, the use of data analytics techniques, as well as using machine learning, artificial intelligence and cloud-based warning systems; for example a system that indicates an airline company when to reschedule flights to avoid thunderstorms or a farmer when to irrigate a particular row of crops. These models can be programmed to predict how the atmosphere and the weather will change. Despite these advances, weather forecasts are still often incorrect. Weather is extremely difficult to predict because it is a complex and chaotic system.

Weather forecasting contributes to the social and economic welfare in many sections of the society. Weather forecast provides vital information to a wide range of fields: agriculture, aviation, energy, commerce, marine, advisories, insurance companies, etc. It can also significantly influence decision and policymaking, global food security, construction planning, productivity and environmental risk management (Wiston, 2018).

Weather forecasting is often used to predict and warn about natural disasters that are caused by abrupt change in climatic conditions. Catalyzed by climate change, extreme weather is an increasing liability to the economy, with approximately 10 weather and climate disasters costing