Intelligence augmentation using big data towards health care

D. G. Ishara Nuwan Karunathilaka

2019

Intelligence augmentation using big data towards health care

A dissertation submitted for the Degree of Master of Computer Science

D. G. Ishara Nuwan Karunathilaka

University of Colombo School of Computing

2019

Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: D.G. Ishara Nuwan Karunathilaka

Registration Number: 2016 MCS 052

Index Number: 16440521

_____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr./Ms.

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. Ajantha Athukorala

_____

Signature: Date:

Acknowledgement

This thesis is the result of me being fortunate to have the unconditional assistance of several people who have been extremely supportive in various ways. First and foremost, I would like offer my humble gratitude to Dr. Ajantha Athukorala, of the University of Colombo school of computing. My supervisor, for the tremendous encouragement, support and the guidance given throughout this research.

My sincere thanks goes to all the lecturers at the University Of Colombo School Of Computing for their valuable advices and comments given at various stages of this research.

I thank my all friends who were there around me, with best of their encouragement, suggestions and support throughout this research.

Last but not the least, I would like to thank my family my parents, for giving me birth to me at the first place and supporting me throughout given life.

Abstract

Artificial intelligence (AI) is leading technology hype in the world. Because of the machine performance, low hardware cost and growth of the big data, every industry willing to have artificial intelligence-based product and services to their end customers. Rise of the artificial intelligence also make bad impact on human job security.

Intelligence augmentation other hand, it closely works with humans and helps humans to make better decisions without replace them.

This project aims to identify intelligence augmentation solution to big data available in health care industry with the help of supervised machine learning method.

Key words

Artificial Intelligence, Intelligence augmentation, Machine learning, Big data, Supervised learning

Table of Contents

List of Figures

Chapter 1

1.1 Introduction

As Technology grows, data associated with each industry grows rapidly. Each industry processes this data to get more insights of their business and make strategic business moves. Low cost of community hardware and growth in cloud computing has led to processing these large amounts of data using new data processing technology to derive insights and transform the way people live, think and work

Artificial Intelligence is one of the fast-growing technology areas that leverage deep learning to train machine to learn from experience and adjust to new inputs and perform human like tasks. Big data helps AI to build intelligence, the combination of these two factors gives the ability for machines do tasks like humans. There are pros and cons with regards to AI There is a fear building around AI where some think that human jobs will be replaced by Robots. As a solution propose Intelligence Augmentation or IA, that aims to assist humans over similar machine learning techniques rather than replace humans. In Intelligence Augmentation important decisions will be taken with the intervention of humans without replacing them.

1.2 Statement of the problem

Health care industry is growing with the aid of digital technology. Smart devices and wearable generate large amounts of digital data on a daily basis. This digitization of data leads to the need of artificial intelligence solution for data analysis. Health care is a more sensitive area for decision making and humans tend to be the better decision makers with the help of machines for data analysis. Medical errors are the third leading cause of death in the world, because of the cognitive errors. Hence giving control of discussion making to machines in the health care sector is risky. Instead, AI can be used to reduce cognitive workloads for physicians.

This research focuses on how Intelligence Augmentation (IA) can be used in the healthcare industry and how smart humans and smart machines can core exist and create a better outcome than working in silo

1.3 Aims and Objectives

Project Aim:

Build an Intelligence Augmentation dash board to made batter decisions referring to the past data and redactions proposed by the machine learning models.
Project Objective:
In order to attain the above project aim, the following objectives are to be achieved initially.

- Collect relevant data from online and prepare comprehensive data set.
- Extensive search of the literature and finalize the machine learning model that will use for the predictions.
- Select the appropriate testing and evaluation strategy for the model

- Build the relevant model and augmentation dash board using python language.

1.4 Scope and Limitation

- Publicly available data and machine learning algorithm will be used in this research
- Data preprocessing will apply for data consistency. That will be achieved by python libraries.
- This project will not cover every disease and every sub area of the health care. Thus, the solution in this project aims to introduce a reference framework that guides IA solution to every area in the healthcare industry.

1.5 Thesis Outline

The thesis organized as follows.

Chapter 2, Literature Review describe the current research about Intelligence Augmentation over big data towards healthcare.

Chapter 3, Methodology describes regarding the selection of the experimental environment through tracing and analysis. Therefore, this section of the research describes the proposed approach followed by the experimental platform from which the data has been collected, as well as the overview of the dataset that is collected and analyzing methodology. And also, it explains algorithm used in research and testing methodology.

Chapter 4, Analysis & Results presents experimental set-up, test observations and results acquired by algorithm's executions. Also, explore the final augmentation dashboard.

Chapter 5, Conclusion and Future Works summarizes research and highlights the new directions expected for future works, where more efforts should be taken with the aim of enhancing the accuracy and efficiency of this research.

Chapter 2

2.1 Literature review

Healthcare industry is vastly digitization industry. There are lot of data generating from equipment's, wearable and IOT devices. Those data provide good background for AI based solutions for Healthcare industry.AI is the hot buzz world and trend in the IT industry. Because of the success of AI based applications most industries willing to build AI based solutions for their problems.

Artificial Intelligence (AI) think like humans and mimic the way a person acts, fear around the people that they will be replaced by the machines. In health care industry jobs roles like radiologists, cardiologists, oncologists are at risk of replacing with AI based solutions. This study will focus on how IA (Intelligence augmentation) help to make better decisions in healthcare industry without replace human with machines.

2.1 Intelligence Augmentation (IA)

Intelligence Augmentation is an alternative to the AI concept. That is focus on AI assistive role design to enhance the human intelligence rather than replace them. Augmented intelligence, which has more natural connection than the artificial intelligence, will help people to understand that augmented intelligence will simply improve product and services, not replace the humans that use them

Intelligence augmentation also assist the accuracy of current state of technology and research. Artificial intelligence programs making decisions after analyzing the patterns in large data set. good decision of artificial intelligence programs depend on the data that human input to the program. The word augment means “to improve”, reinforces the role of human intelligence plays when using machine learning and deep learning to discover patterns in data sets.

Intelligence augmentation all about empowering humans with tools that are more helpful them to make better decisions.

2.2 Bid data

Data is the most important factor for any solution. Quality of data drive to the good solution. Because of the digital age large amount of data generating every day. Big data refers to massive volume of both structured and unstructured data beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety.

Data is the most important thing in the AI. That will drive to build accurate and quality machine learning model. Quality of the data cause for the quality and accuracy of the model. Both structured and unstructured data can be used to create machine learning models. So big data is most important requirement when build and maintaining the machine learning and AI based solution.

2.3 Supervised learning vs Unsupervised Learning

Machine learning is essential part of modern data science and the enterprises. Machine learning themselves and improved themselves. In world today most of the tech and product companies use machine learning techniques to improve their product and services. Machine Learning mainly have three types supervised learning unsupervised learning and feedforward learning.

Supervised machine learning method is the commonly use machine learning technique. In supervised learning data is labeled and structured. Machine is train for get label output for labeled input. After training with label input machine can predict the out for new input data. Supervised learning can be divided into two categories as classification and regression. Classification process of input into desirable output "true" ,"false" or "yes", "no". Regression process of input to predict real value as "what would be the price of house with 10 perch in Colombo 7?"

Unsupervised learning doesn't require labeled input all input data set are unlabeled. Unsupervised learning tries to group data into similar groups based on the similarities. Clustering is the main technique used in unsupervised learning. Clustering able to process and identifying inherent groups inside the input data.

In this project supervised learning used for big data available in health care industry to create a model, so human can augment that model to get better decisions with high accuracy.

2.4 Related Work

"Augmented Intelligence: Enhancing Human Capabilities" [1] this paper explore the how augmented intelligence help in the Digital assistant. On this research they consider speech as input and voice output after applying the augmented intelligence. This research is not focus on the area of health of healthcare and didn't provide augmented solutions for that area.

"Designing Intelligence Augmentation System with a Semiotic-Oriented Software Development Process" [2] this paper discusses about how computer software makes intelligence decision making collaboration with ongoing discussion between human user and the computer system. Final decision making is not provided neither by the human being nor by the computer but is collaboration of the two of them. This research also not focus on the healthcare but this one focus on the intelligence augmentation area. Propose research will focus on how intelligence augmentation apply for a healthcare. Using machine learning first identified the patterns and help of intelligence augmentation make the final decision.

"Data augmentation importance for classification of skin lesions via deep learning" [3] This research paper discussed the how Data augmentation is important to build a classifier for type "Melanoma", is a fatal type of cancer. This research mainly focusses on the input data to the machine learning algorithms. How augmented intelligence use for identifying the correct images as input data. Propose research focus on the how augmented intelligence use for making final decision collaboration with machine learning.

“Data augmentation using synthesized images for object detection” [4] this paper discuss about generate dataset for deep learning algorithms by synthesizing the image of background and object. This is also a data augmentation research not specific for the healthcare industry.

“Data Augmentation for deep neural network acoustic modeling” [5] this research paper focus on data augmentation approaches to acoustic modeling using deep neural networks (DNNs) for automatic speech recognition (ASR). propose research will differ from this on is, proposed on mainly focus on data and not for voice. Also propose one working with data of the healthcare and collaboration of augmentation intelligence make the final decision.

As a conclusion above all previous researches focus on the augmentation intelligence for different aspects. Not specifically for the healthcare. Propose research is focus on the data in healthcare and collaboration of machine learning and augmentation intelligence to make a final decision.

Chapter 3

3.1 Methodology

Methodology of this research is beginning from selecting of the data set and methodology of the data analysis and finally build an augmentation dashboard. This section describes the proposed approach for followed. The general view of the process graphically shows in the Figure 3-1.

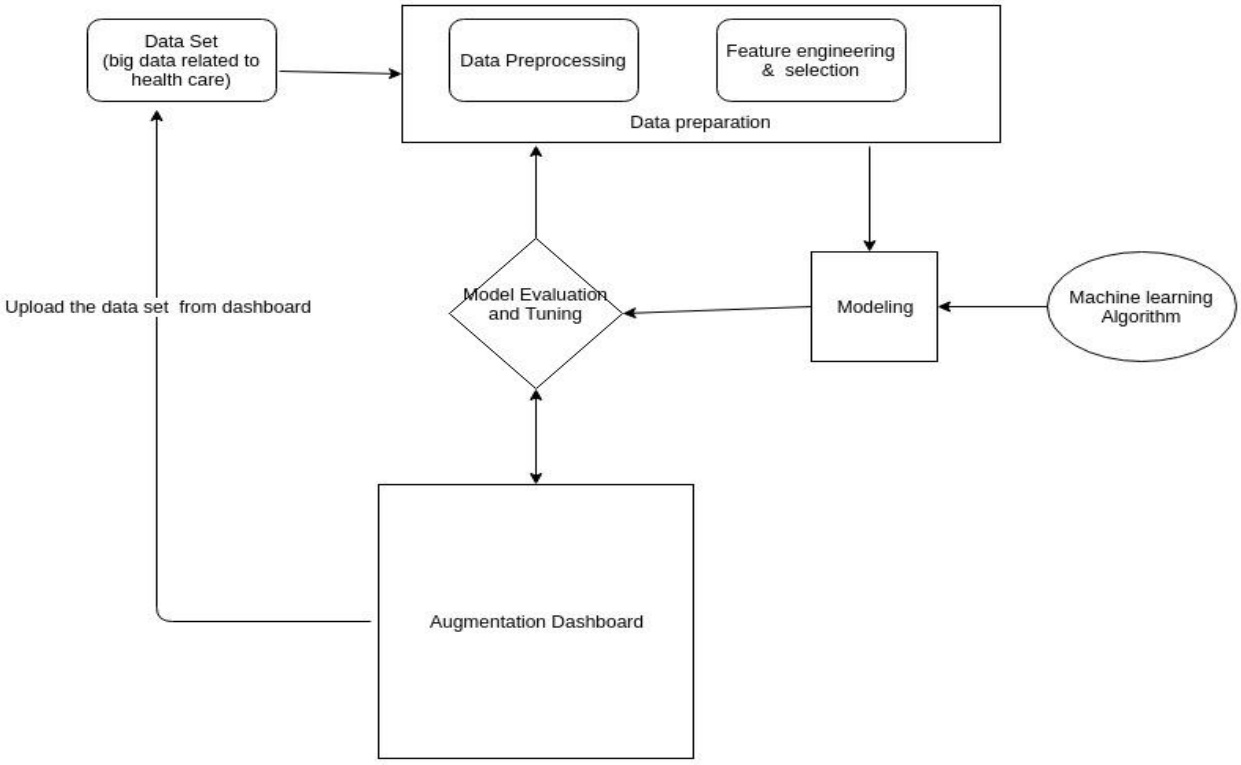

Figure 3-1 General process

3.2 Data collection

Publicly available data set choose for this project. This data set has 33 columns and 100000 rows with .csv format.

3.3 Data preparation and pre-processing

In this stage all redundant data are removed and all unnecessary data columns which are no required to model creation are removed. Also, empty data rows also removed. This data removal part is not manual and not directly apply to the csv file. This is done by the python language and it is happening on machine memory not in the csv file.

3.4 Feature selection

Correct features are identified after drawing the correlation matrix. Using this Metrix can identify the positively and negatively correlated features. Without duplicating feature, we can omit the correlation feature.

3.5 Algorithm Selection

As per mentioned earlier in the literature review chapter. After researching about many supervised learning algorithms and related work decide to implement the system with the help of Random forest algorithm

3.6 Augmentation dashboard

This dash board build using python, html and CSS. So, all the visualization of data and machine learning model information and capture input data for prediction and model predictions visualize in the dashboard.

Chapter 4

4.1 Evaluation & Result

Evaluation approach for this project will be mainly opinion and interview based. So even before the augmentation dashboard deployed, there will be questionnaire or survey to selected and experts in the healthcare industry who use medical data in day to day life. The questionnaire/ survey would contain, core questions such as basic and important information they expect from the dashboard, how would be the dashboard look like. The result of these questions would provide some information's which could be utilized to enhance the dashboard or go back and revert the any functionality or models which aren't useful.

4.2 Tools selection

There are mainly R language and python language can be used to implement the research. Python and PyCharm are used to data analysis algorithm and dashboard implementation.

Python is an interpreted general-purpose programing language. It can run on any operating system and it most popular language than R. Also it is widely used among the data scientist and data miners.

4.3 Main reason to choose python

- Data wrangling:

Process of cleaning messy and complex data sets to convenient analysis called data wrangling. python support most of the data formats and have libraries to data manipulate and wrangling.

- Data visualization

Data visualization is one of the important aspect of data analysis. Python can visualize most complex visualizations correctly using libraries.

- Availability

Python is open source and have huge community support and libraries to do the data science and data mining tasks.

4.4 Testing Methodology

The initial step of the testing process involving creating two set of data, a train data set and the testing data set. Training data set use to train the model and test data set used to validate the model. For this research 70% of the data compromise to train the model and rest of the 30% was selected to test the model.

4.5 Result

After all the machine learning process, random forest classifier shows the model accuracy as 0.9532

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import accuracy_score

# split data train 70 % and test 30 %
x_train, x_test, y_train, y_test = train_test_split(x_1, y, test_size=0.3, random_state=42)

#random forest classifier with n_estimators=10 (default)
clf_rf = RandomForestClassifier(random_state=43)
clr_rf = clf_rf.fit(x_train,y_train)

ac = accuracy_score(y_test,clf_rf.predict(x_test))
print('Accuracy is: ',ac)
cm = confusion_matrix(y_test,clf_rf.predict(x_test))
sns.heatmap(cm,annot=True,fmt="d")
```

```
/home/ishara/anaconda3/envs/isharaEnv/lib/python3.7/site-packages/sklearn/ensemble/forest.py:24
ault value of n_estimators will change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
Accuracy is:  0.9532163742690059
```

Figure 4-1 Model accuracy

Chapter 5

5.1 Conclusion & Future works

Intelligence augmentation is the future of machine learning both humans and machines work together to solve complex problems with high accuracy. Every domain need intelligence augmentation but healthcare industry is the most sensitive area because it is dealing with the human life. Early predictions, detections and analysis of diseases leverage human life.so Intelligence augmentation is most prominent requirement in healthcare industry.

This research present augmentation dashboard with cancer prediction model. Where you can upload data set to the dashboard, background it will clean the data, do the feature engineering and create a model using random forest algorithm based on the inputs data from the dashboard do the predictions and visualizations from background and update the dashboard accordingly

5.2 Future work

In this current research data obtained from freely available source in the internet. Future is better to find real data from relevant parties and use as data set to the machine learning model.

Currently this dashboard and machine learning model support only one decides. This can be extended to multiple devices and make this dashboard as central augmentation source.

Currently this model is build using random forest algorithm. Future there will be new algorithm or new techniques to tweak the model for higher accuracy rate model.

References

[1] Akshay Hebbar,"Augmented intelligence: Enhancing human capabilities" in 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) - 2017

[2]Andre Paraense,Ricardo Gudwin and Rodrigo Goncalves, "Designing Intelligence Augmentation System with a Semiotic-Oriented Software Development Process" in Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007) - 2007

[3]Enes Ayan, "Data augmentation importance for classification of skin lesions via deep learning", in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) -2018

[4]HyunJun Jo, "Data augmentation using synthesized images for object detection", in 2017 17th International Conference on Control, Automation and Systems (ICCAS) - 2017

[5]Xiaodong Cui, ”Data Augmentation for deep neural network acoustic modeling”, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) - 2014