Big Data Analytics for Healthcare Services Using C4.5 Algorithm on Map Reduce

Pau Suan Mung, Sabai Phyu
University of Computer Studies, Yangon
pausuanmung@ucsy.edu.mm, sabaiphyu@ucsy.edu.mm

Abstract

Healthcare industry is an ever-increasing rise in the large amount of records such as doctors, patients, medicines and medical history. Although previous medical records are beneficial for not only individual but also human society, maintaining and analyzing large amount of such data is a big problem. Traditional data mining tools are inadequate for such amount of data. Big data analysis can be used in various applications with different domains like education, security and health care. This system uses MapReduce based C4.5 Decision Tree algorithm for health care big data to manage, analyze and extract the most suitable data for right conditions. The classification rules produced by this system can be used to classify the particular disease. Tuberculosis disease is used as a case study in the system.

Keywords*: Big data, Healthcare data, Machine learning, Classification*

1. Introduction

The enormous amount of data generated everyday by industries such as education, finance, business, manufacturing, healthcare is referred to as Big Data. The traditional storage such as DBMS and RDBMS are incapable of storing and traditional data mining techniques are also inadequate for analyzing on Big Data. Developing effective tools and technologies is a need. If Big Data is used in helping to make the world a better place, there's no better example than it is used in healthcare. [3] The healthcare industry can be seen as highly data intensive. In healthcare industry, data is a collection of patient data and they are large, distributed, complex and growing so fast. Maintaining and analyzing traditional tools and technologies on such large amount of Big Data are insufficient. If Big Data is managed well, it can be used to unlock new sources for economic value and scientific discoveries. It can assist in policy making on Government level. In the field of Health Informatics, researchers discussed a number of big data analysis tools and approaches for analysis of healthcare data. In developing countries, 99% of all maternal deaths fall out and around 830 women die every day from preventable causes. In low and middle-income countries, over 95% of TB deaths occur, and one in three of HIV deaths is due to TB. Health specialists said both diseases remain a serious public health threat in such countries and called for an intensified public awareness campaign to fight them [8]. According to this large amount of statistic, Big Data system can be built to maintain healthcare records from different sources of different locations. These records will be analyzed on different nodes using Hadoop and Map Reduce framework and then clustering and classification will be applied to get the specific output. Some of the data sources are US government open data at https://www.data.gov, https://www.cdc.gov and research data source from https://archive.ics.uci.edu/ml/datasets.html [3].

2. Related Works

In the paper [6], the authors proposed a framework for Big Data Analysis on healthcare data. The real healthcare data from the data warehouse of Vancouver Island HEALTH AUTHORITY (VIHA) was used. Big data analysis tools such as MapReduce and Hadoop (HDFS) were used with HBase. Apache Phoenix was used to satisfy the query performance.

Authors in paper [13] presented big data recommendation system using Naïve Bayesian classification on Apache Mahout. This system assisted to recommend user's health conditions. When a disease was identified, the right care is delivered with necessary treatments suggestions. Database query was made through Hive command.

In the paper [11], the authors proposed useful big data tool for healthcare analysis. It used k-mean clustering to get right data for right patients at right time for right living and care. This system used

MongoDB for storage of big data and healthcare data from around the world were extracted.

The paper [14] proposed efficient use of big data in public health proposes. Data from San Diego County of California was collected and much of variables related to healthcare were analyzed. Disease prevention and health promotion for public health are emphasized in communities than individuals.

There are many other authors who emphasize on big data in healthcare analysis. In the paper [7], it presented the survey of analytical processes of big data on healthcare data. It focused on classification and clustering. Detail of decision tree and support vector machine classification were presented. The other paper [3] included big data analytics in healthcare, its benefits, phases and challenges. It also presented the 5V of big data characteristics, the phases in the big data process and the benefits of using big data analytics in healthcare. It showed the use of healthcare big data from point of view of individuals, hospitals, doctors and even governments. It concluded with the challenges found in healthcare big data analytics.

3. Big Data

Big data is to a collection of data sets with large, complex and difficult to process with traditional data processing. Big data is collected from different sources and such data may be structured, semi-structured and unstructured [15]. There are many challenges of big data. As data in big data is big, doubling data volume emerges in size about every two years. Organizations using big data must still struggle to keep pace with their data and find ways to effectively store it. Big data technology is changing at a rapid speed and thus keeping up with big data technology is an ongoing challenge [9].

Big data analytics are used on big amounts of data and uncover correlations between data, some hidden patterns in big amount of data and other insights. Some of biggest players:

- ***Data management***: Data must be high quality and well-governed before it is reliably analyzed.
- ***Data mining***: It helps examine big amounts of data to discover patterns in the data [12].

Several types of technology work together because there's no single technology for big data analytics. Some are: Apache Hadoop, Hadoop Distributed File Systems (HDFS) and MapReduce framework.

Apache Hadoop is an open source framework and consists of two main components: distributed processing framework named MapReduce and distributed file system called Hadoop Distributed File System (HDFS).

Hadoop Distributed File System (HDFS) is the storage component and it provides distributed architecture for large scale storage. When a file is stored in HDFS, the file is divided into evenly sized blocks.

MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster. It works by breaking the processing into two phases: the Map phase and the Reduce phase. Each phase has key-value pairs as input and output. It also specifies two functions, Map and Reduce functions. The output from the map function is processed by the MapReduce framework before being sent to the reduce function.

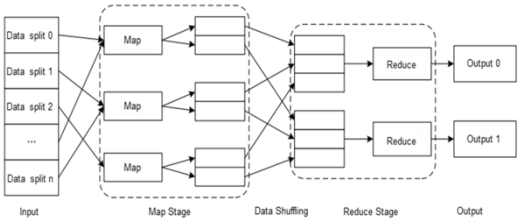

Figure 1. MapReduce Programming Model

4. Classification with Decision Tree

Classification is used to classify the input data with the target classes that are used by classifier in training and testing process [7]. Input data is set into machine learning algorithm. After the classifier is trained using the training data, next phase is testing in which testing data is given to perform the accuracy of the system. New data can be applied for prediction about target data [6].

Decision tree is the form of tree structure and is used for classification or regression models. Decision tree breaks down a dataset into smaller and smaller subsets while an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or a decision. The topmost decision node in a tree which corresponds to the best predictor called root node [5].

Algorithm: C4.5 Decision Tree

Input : training dataset *T*, attributes *S*,
Output : decision tree, *Tree*

1. **if** T is NULL **then**
2. **return** failure
3. **end if**
4. **if** S is NULL **then**
5. **return** Tree as a single node with most frequent class label in T
6. **end if**
7. **if** |S| = 1 **then**
8. **return** Tree as a single node S
9. **end if**
10. **set** *Tree* = {}
11. **for** s ∈ S **do**
12. **set** *Info*(s, T) = 0 and *SplitInfo*(s, T) = 0
13. **compute** *Entropy* (s)
14. **for** v ∈ values(s, T) **do**
15. **set** $T_{s,v}$ as the subset of T with attribute s = v
16. *Info*(s, T) += $\frac{|T_{s,v}|}{|T_s|}$ *Entropy*(s_v)
17. *SplitInfo*(s, T) += - $\frac{|T_{s,v}|}{|T_s|}$ log $\frac{|T_{s,v}|}{|T_s|}$
18. **end for**
19. *Gain*(s, T) = *Entropy*(s) – *Info*(s, T)
20. *GainRatio*(s,T) = $\frac{Gain(s,T)}{SplitInfo(s,T)}$
21. **end for**
22. **set** s_{best} = argmax{GainRatio(s, T)}
23. **attach** s_{best} into *Tree*

Figure 2. C 4.5 Decision Tree

C 4.5, as shown in Figure 2, is a model algorithm for decision tree, in which information gain ratio is used as the splitting point.

$$\text{Entropy (S)} = -\sum_{j=1}^{C} P(S,j) * (S,j) \qquad (1)$$

Equation 1 is to find the ratio of instances in S with j^{th} class label, and total number of classes, C.

$$\text{Info (S, T)} = -\sum_{v \in V}^{C} \frac{|T_{S,v}|}{T_S} \text{Entropy } (S_v) \qquad (2)$$

Equation 2 is to find the information gain needed after splitting by attribute S, in which T_S is the subset of T by attributes S, and T_S, v is the subset of T_S of value v for attribute S, and Values (T_S) is the set of values for attribute S for records in T_S. The Equation 3 is used for SplitInformation.

$$\text{Split Info (S, T)} = -\sum_{v \in V}^{C} \frac{|T_{S,v}|}{T_S} * \log \frac{|T_{S,v}|}{T_S} \qquad (3)$$

The information gain ratio is:

$$\text{Gain Ratio (S, T)} = \frac{Gain(S,T)}{SplitInfo(S,T)} \qquad (4)$$

The attribute with the maximum Gain Ratio is used as the splitting attribute.

C4.5 Decision Tree works in two steps for classification task: learning and classification. In learning task, the attribute with highest information gain is selected from training dataset. It becomes root and similarly other attributes are also selected.

5. Design of the System

This system is built as a healthcare big data analytics system. In this section, we present a design of the framework of the system as shown in Figure 3.

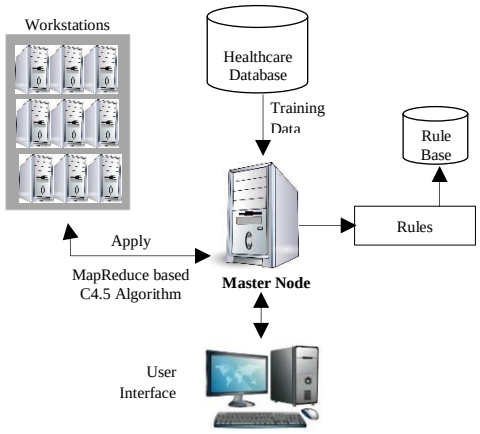

Figure 3. Framework of the System

The big data about health care stored in database is loaded to the HDFS and these dataset is used as input for the algorithm. Then MapReduce based C4.5 decision tree algorithm is applied on this data. There are three steps in MapReduce based C4.5 algorithm namely attribute selection, dataset splitting and tree building.

Attribute Selection: The main task of decision tree construction is the ratio of information gain for selecting attributes. An attribute with maximum information gain ratio is selected as the root and also the splitting attribute. As the traditional way, this step of the system chooses the splitting attribute with highest information gain. Although the information gain is calculated with the whole dataset in traditional method, this system finds the information based on several small datasets that are located on each node. The sum of information gain of a particular attribute in each node is used to find the splitting attribute, a_k, with highest information gain. The process of this stage is shown in Figure 4.

Dataset Splitting: After the best splitting attribute a_k and the cut points cp_k are identified, the dataset is split into several subsets. The process of dataset splitting is shown in the algorithm of Figure 5. If the largest class probability, max{P_k}, in X_{id} is bigger than $\theta \in [0,1]$, X_{id} is created as a leaf node and this node is not included in the next splitting.

Algorithm: Attribute Selection
Input : A training dataset X
Output : Splitting attribute a_k and cut points cp_k

In the j[th] Mapper
Input : X_i of X, i=1,2,…,N
Output : <key, value> = <a_k, [Ratio, cp_k]>
1. **for** each attribute a_k, k=1,2,…,n do
2. Sort its values x_{1k}, x_{2k}, …, x_{Nk}
3. Find all cut point cp_{ik}, i=1,2,…,N-1
4. **for** each cut point cp_{ik} do
5. Calculate InfoGain Using Equation 2
6. **end**
7. Select optimal cut point cp_k
8. Calculate the splitInfo for cp_k using equation 3
9. Calculate the GainRatio using equation 4
10. ***Mapper Output:***
<key, value> = <ak, [Ratio, cp_k]>
11. **end**

In the Reducer
Input : <key, value> = <a_k, LIST[Ratio, cp_k]>
Output : <key, value> = <a_k, BEST[Ratio, cp_k]>
1. **for** each attribute a_k do
2. Ratio = $\sum_{j=1}^{m} Ratio_j$
3. $cp_k = \frac{\sum_{j=1}^{m} cp_j}{m}$
4. **end**
5. ***Reducer Output***:
<key, value> = <a_k, BEST[Ratio, cp_k]>

Figure 4. Attribute Selection Algorithm

Algorithm: Dataset Splitting
Input : A training data X,
The splitting attribute a_k
The cut points cp_k and empty set Q.
Output : Dataset Q

In the j[th] Mapper
Input : X_j = {x_i, i=1,2,…,n}
x_i is the i[th] instance with n attributes
Output : <key, value> = <id, x_i>
where id is label of output subset
1. **for** each instance x_i of X, i=1,2,…,N **do**
2. **if** x_{ik} > cp_k **then** id = 1
3. **else** id = 0
4. **end**
5. ***Mapper Output:***
<key, value> = <id, x_i>
6. **end**

In the Reducer
Input : <key, value> = <id, LIST[x_i] as X_{id}>
Output : <key, value> = <id, X_{id}>
1. **for** each X_{id} **do**
2. Define K is the class number and P_k is the class probability
3. **if** K > 1 and max{P_k} < θ **then**
4. Build a no-leaf node
5. Add X_{id} to Q
6. **else**
7. Build a leaf node
8. **end**
9. ***Reducer Output:***
<key, value> = <id, X_{id}>
10. **end**

Figure 5. Dataset Splitting Algorithm

Tree Building: In the tree building step shown in Figure 6, node selection is essential. Considering which attribute is most suitable in a particular node is the most important. The attribute selection algorithm shown in Figure 4 is used to get the best splitting attribute. Then dataset is split into several nodes with the help of dataset splitting algorithm shown in Figure 5.

Algorithm: Tree Building
Input : A training dataset, D_0
Output : An Decision Tree
1. D is initialized as an empty set
2. Select one node, M, from D_0
3. **for** each M of D_0 **do**
4. Get the best splitting attribute a_k and cut points cp_k by Attribute Selection Algorithm
5. Split M into n child nodes {m_i, i=1,2,…,n} based on a_k and cp_k by Dataset Splitting Algorithm
6. Add {m_i, i=1,2,…,n} to D.
7. **End**
8. TreeDepth++.
9. **if** D is not empty **then**
10. D_0 = D
11. Recursive search the new node from D_0 by Tree Building Algorithm
12. **end**

Figure 6. Tree Building Algorithm

6. Experimental Results

In this experiment, the records of patients of Tuberculosis (TB) disease are used as case study. The attributes and descriptions of this disease are as shown in Table 1. Three types of TB diseases, Pulmonary TB, Extra Pulmonary TB and No-TB, are used as class labels.

The decision tree generated from algorithms is shown in Figure 7 and the rules generated from this decision tree is shown in Figure 8.

Figure 9 shows comparison of runtime between original C4.5 algorithm and MapReduce Based C 4.5 algorithm. Although there are a little different between small amount of data, this different is larger for large amount of dataset. Original C4.5 algorithm is tested with Weka machine learning tool and MapReduce based C4.5 algorithm is tested on Map Reduce with three nodes. The more the dataset, the more the efficient obtained.

Table 1. Attribute Description for TB Disease

Attributes	Description
Duration of illness	Length of illness over 3 weeks (Binary)
Expectoration	Abnormal sputum production (Binary)
Haemoptysis	Splitting of blood from the lungs (Binary)
Chest pain	Pain at the chest (Binary)
Breathlessness	Shortness of breath (Binary)
Weight Loss	Loss of body weight (Binary)
Low Grade Fever	Fever with low temperature (Binary)
Night Sweating	Sweating at night (Binary)
Loss of Appetite	Decrease appetite (Binary)
Malaise	Feeling of general discomfort, uneasiness or pain (Binary)
Easy Fatigability	Reduced capacity to maintain activity (Binary)
Tuberculosis Intrathoracic Lymph	DX-NAD/ DX-KOCH
Irritability	Feeling of agitation (Binary)
Pre-Treatment Smear	Sputum Examination (3+/2+/1+/Neg)
Family History of TB	Binary
Class	Pulmonary TB (P), Extra Pulmonary TB (EP), No-TB

7. Conclusion and Future Works

In this paper, we have presented the framework for a big data analysis for healthcare data and MapReduce based C4.5 algorithm. In this system, Hadoop is used for big data analytics. In each of the node, MapReduce based C4.5 decision tree classification is applied. Healthcare data is collected from different sources such as many hospitals and the whole dataset is divided into two parts, training and testing dataset. Training dataset is used to train the system with the help of decision tree induction and the result of this system is classification rules and these rules are tested with testing dataset. Healthcare data is very important and need a careful analysis because we can learn from previous data to get healthier system. The main aim of this system is to use technology effectively in health sector to save many lives.

There are many limitations and awareness of using healthcare data. Privacy is the most important consideration. Protection previous healthcare data is still a big problem. Most patients do not want to give their data, they want to keep it secretly.

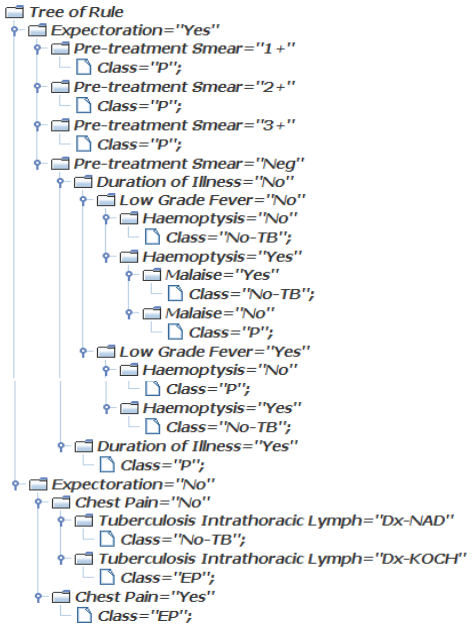

Figure 7. Decision Tree for TB Disease

In this system, we use Hadoop framework for big data analysis. Another big data analysis framework such as Apache Spark can be used. And also, there are many classification methods in machine learning. Using other classification methods and generating comparison of the result is also a future work.

1. IF Expectoration='Yes' AND Pre-treatment Smear='1+' THEN Class="P";
2. IF Expectoration='Yes' AND Pre-treatment Smear='2+' THEN Class="P";
3. IF Expectoration='Yes' AND Pre-treatment Smear='3+' THEN Class="P";
4. IF Expectoration='Yes' AND Pre-treatment Smear='Neg' AND Duration of Illness='No' AND Low Grade Fever='No' AND Haemoptysis='No' THEN Class="No-TB";
5. IF Expectoration='Yes' AND Pre-treatment Smear='Neg' AND Duration of Illness='No' AND Low Grade Fever='No' AND Haemoptysis='Yes' AND Malaise='Yes' THEN Class="No-TB";
6. IF Expectoration='Yes' AND Pre-treatment Smear='Neg' AND Duration of Illness='No' AND Low Grade Fever='No' AND Haemoptysis='Yes' AND Malaise='No' THEN Class="P";

Figure 8. Rules Generated from the Decision Tree

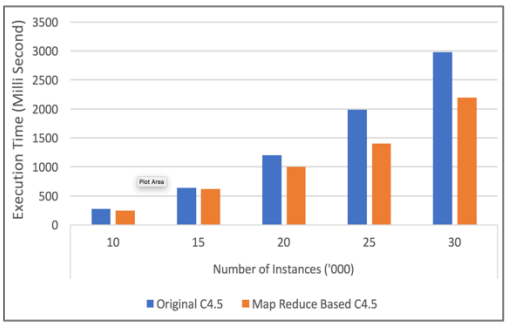

Figure 9. Runtime Vs. Input Size

References

[1] Bernard Marr, "How Big Data is Changing Healthcare",http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/

[2] Dumbill, E., "What is Big Data?", https://www. oreilly.com/ideas/what-is-big-data, 2012

[3] Jasleen Kaur Nains, "Big Data Analytics in Healthcare – Its Benefits, Phases and Challenges", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 6, Issue. 4, 2016

[4] Luke Dormehl, "President Obama's New Healthcare Initiative will Harness the Power of BigData", http://www.fastcompany.com/3041775/fast-feed/president-obamas-new-healthcare-initiative-will-harness-the-power-of-big-data

[5] Machine Learning 101, "Decision Tree Classifier – Theory", https://medium.com/machine-learning-101/chapter-3-decision-trees-theory-e7398adac567

[6] Mu-Hsing Kuo, Dillon Chrimes, Belaid Moa and Wei Hu, "Design and Construction of a Big Data Analystics Framework for Health Applications", IEEE ICSC, 2015.

[7] Muhammad Umer Sarwar, "A Surver of Big Data Analytics in Healthcare," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 8, No. 6, 2017.

[8] Myanmar Times, https://www.mmtimes.com/news/tb-still-serious-threat-say-doctors.html, 2018

[9] Oracle, "What is big data? The value of big data.", https://www.oracle.com/big-data/guide/what-is-big-data.html

[10] Paola Galdi, Roberto Tagliaferri, "Data Mining: Accuracy and Error Measures for Classification and Prediction", https://www.researchgate.net/ publication/322179244_Data_Mining_Accuracy_and_Error_Measures_for_Classification_and_Prediction

[11] Priyanka Dhaka and Rahual Johari, "HCAB: HealthCare Analysis and Data Archival using Big Data Tool", IEEE, 2016.

[12] SAS, The Power to Know, "Big Data Analytics: What it is and why it matters", https://www.sas.com/en_ph/insights/analytics/big-data-analytics.html

[13] Weider D. Yu, Choudhury Praktiksha, "A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment", IEEE 39th Annual ICSA, 2015.

[14] Yannis Katsis, Batasha Balac, "Big Data Techniques for Public Health: A Case Study", IEEE/ACM, ICHASE, 2017.

[15] Zikopoulos, P. and Eaton, C., "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill Osborne Media, 2011.