A Review Paper on Big Data in Lung Cancer
Big Data Analytics in Lung Cancer

Ritu Parna Panda, Prakalpa Prakash Barik and P. Alok Kumar Prusty
Computer Science Department, Centurion University of Technology and Management, Bhubaneswar, India

Abstract-- Big data is a declaration used to recognize the database whose area is afar the potential of typical database software tools to store, organize and examine. Big data has shown a new path towards the mankind. Cancer is a tumor of diseases involving abnormal cell growth with the possibility to occupy or disperse to other parts of the body. It has been a dead line for millions of people. Every cancer is particular and the medical treatment is complex. The data of each cancer patient is too large and it varies from one person to person.

This position paper will traverse censorious tools and tackle that accelerate knowledge locating along the cancer study continuity. Let’s discuss something more about this topic in this paper.

Keywords-- Bigdata, Lung cancer, Methodology

I. INTRODUCTION

Cancer is a miscellaneous disease. It is also familiar that a variety is present within a solitary patient. It is clear that actual alteration

Collect mutually unlike unaccompanied cancer cells. Our existing mechanization for cancer biomarker finding is normally proficient of identifying a snapshot, but not the energetic and long-term modification of the cancer landscape. Big data can block this barrier by assembling dissimilar molecular characteristic at the DNA, RNA, and protein and metabolite proportion. Bio data-mining has guided those working in the existence sciences to accept the information area, in order to assist the barrage of big data produced by next-peer biotechnologies, such as next-peer sequencing, proteomics and metabolomics, as well as the arranged and disarranged medical and healthcare statistics from computerized well-being evidence.

Figure 1:
Big data definitely throws dare on cancer research in different condition. The trends of organizing different platforms and their findings create a junction among different sectors and measurement, which is ultimately a step towards an entire understanding of the disease. Nowadays, next-generation syncing, paramedics and meta bolomics develop fast in different directions. However, they also create a difficult question regarding the existing platform of using a single feature to measure cancer. Multifaceted approach to cancer research is urgently needed.

On the other hand, there is a multitude of data, not only huge in volume and complicated in structure, but also vast in dynamic scale and depth. We for sure that a large amount of loosely connected, naturally noisy and miscellaneous data may also be collected in the databases. All data do not contain useful data, we cannot judge whether the information is useful or junk. Thus, certain standards, protocols or extension may be needed to develop useful data in the same database. Some international corporation and institutions have made achievement to set guidelines to guarantee accuracy and endurance so as to authorize the suitability of the data.

Electronic health records are becoming a better resource for databases. The information can be change and together via telemedicine and mobile connectivity. Patients’ information and data security become an issue and more security part need to be in place. We are in the heed medicine era, and precise decision making based on individual, detailed profiles is a pressing need. Tumor molecular profiling has enabled the sub division of cancer to reconsider treatment system. Big data are exactly sailing in the same boat with correct medicine. Nowadays, more and more databases have been ingrained, e.g., the directory of actual variation in Cancer (COSMIC), which is the largest database of actual variation and their effects on human cancer. It is envisioned that in the years to come, we will step into a remodel era of applying big data to asset to our cancer patients.

II. BIG DATA

Big-data’ is similar to ‘a few-data’, but larger, but having data larger therefore requires different approaches techniques, tools & architectures to fix new problems and old problems in a superior way.

Figure 2:
Key enablers for the growth of Big Data are increase of storage space, boost of processing power and availability of data. Big data’s distributed storage system built on Google File System.
The data’s are stored in rows and columns. It is optimized for sparse, persistent, multidimensional sorted map.

III. LUNG CANCER

Lung cancer is the major source of cancer deaths in the United States and will assert more lives this year than other cancers. Exposure to tobacco smoke is the number one reason people develop lung cancer.

A. Methods Of Testing Lung Cancer-

1. Chest x-ray - A chest x-ray is basically the initial test implemented to figure out any concerns that are found in a careful history and physical. This may display a lump in the bronchi or expanded lymph nodes. Consistently the chest x-ray is usual, and further tests are required to examine for a suspicious lung cancer. Even if a lump is found, we can’t think it as a lung cancer. So further tests are required.

2. CT scan (computerized tomography) - A CT scan is commonly the second step either to follow up on an anomalous chest x-ray finding or to check out annoying symptoms in those with a traditional chest x-ray. CT scanning associated a series of x-rays that create a 3-dimensional vision of the lungs. If the CT is anomalous, the analysis of lung cancer still needs acceptance through a sample of tissue by one of the operation below.

3. MRI (Magnetic Resonance Imaging) - In a few patients, MRI will be used to check out the probability of lung cancer. This operation uses magnetism and does not associated radiation. Certain persons, such as those with metal embed (pacemakers, etc.) should not have MRI scans. The technician will inquire questions to confirm these are not present.

4. PET scan (Positron Emission Tomography) - A PET scan usage radioactive material to create colourful 3-dimensional images of a part of the body. This type of scan different from the others in that it defines tumours that are actively growing. As an expansion to the other operation, a few researchers approved that PET scanning may analyse tumours earlier, even before they are distinguishable untypically through other studies.

5. Sputum cytology- After a lung cancer is imagined based on imaging, a sample of tissue is required to approve the diagnosis and resolve the type of cancer. Sputum cytology is the simple way to do this, but its use is defined to those tumours that enlarge into the airways. Sputum cytology is not always authentic and can miss a few cancer cells.

6. Bronchoscopy- In a bronchoscopy, a lung consultant inserts a tube into the airways to dream up and take a sample of the tumour. This operation is used when the tumour is observe in the large airways and can be attained by the scope. Patients are given anaesthesia during this operation to minimize discomfort.

7. Endobronchial ultrasound- Endobronchial ultrasound is a almost new technique for determining lung cancer. During a bronchoscopy, doctors use an ultrasound inquiry within the airway to check the lungs and space between the lungs (mediastinum.)

B. Methodology Of Big Data Analytics

Step 1: Concept statement
Build need for big data analytics in healthcare established on the “4Vs”.

Step 2: Plan
1. What is the problem being addressed?
2. Why is it important and curious?
3. Why big data analytics way?
4. Backdrop material

Step 3: Methodology
1. Variable selection
2. Outcome & insight
3. ETL and data revolution
4. Propositions
5. Platform/tool collection
6. Conceptual model
7. Data collection
8. Clustering, allocation, Analytic techniques - Association, etc.

Step 4: Deployment
1. Evaluation & validation
2. Testing

IV. CANcer DATA SETS

The data consist of data on 40 lung cancer patients used to compare the effect of two chemotherapy treatment in prolonging durability time.

The data set variables are:

1. SurvivalTime: The duration time in days after the treatment
2. Censored: An indicator variable
   0: complete uncensored case
   1: stands for censored case
3. Med_cond: The general medical condition at analysis of lung cancer on a range of 0 to 100
4. Age: Patient’s age at diagnosis
5. Time_diag_study: Number of days between analysis of lung case and cure
6. TumorType: Tumor types
   Squamous
   Small
   Adeno
   Large
7. Treatment: Two levels of treatment
   Standard treatment
   Experimental treatment
Figure 3:

V. STEPS TO EXTRACT VALUE FROM BIG DATA

A. Step One: Process and Clean Data
It is important to verify your data matches your business objectives. If it does not, there are a lot of questions to address: What are the viable proxies? Are there outliers that need to be taken into report? Does the data consist of bias? Are there removed values? Look for functionalities that will correctly address the disparate needs to clean and action the data. There are a number of methods that can be used to impute, or fill in removed values, such as mean interpolation, Kalman, filter, and ARMA. This step is one of the most important, but may take 70-90 rate of your data analysis project time. The quality of your data will highly affect your test results.

B. Step Two: Analyse and Visualize Data
Analyse the processed data and visually inspect the data for arrangements, trend, and clusters. This is the time to examine relationships and build hypotheses according to your recommendation. The easiest way to complete this operation is with the aid of visualization tools. There are a number of elementary yet powerful visual aids, such as scatter plots, line graphs, stacked bar charts, box-plots, and heat-maps.

C. Step Three: Data Store
You can use different methods to facilitate pattern recognition, counting clustering K-Means, market basket inquiry, hierarchical clustering, Kohonen Self-Organizing maps for visualization, principal component study, factor study, and multi-dimensional scaling. Organizations that leverage and mine their data predictively have a important competitive advantage over their rivals, as they can gain important insights and react quickly to enlarge their business in a approach that was not possible without predictive analytics.

D. Step Four: Build Model
Be sure to have a wide scope of models that provide various perspectives of the data. Some possible models to consider are neural networks, ARIMA, decision trees, regressions, SVM, Naïve Bayes classifier, and discriminant analysis. Every algorithm has its propriety, and it is essential to understand that all models have limitations. There could be more than one model that would work for one issue. Avoid over fitting. Understand not only the probable errors, but also the most deliberate ones, and set guidelines to control against making the most serious of false inferences. Be sure to document and interact the presumptions and results clearly.

E. Step Five: Generate Results and Optimize
Anticipating results are used to authorize objective functions in order to achieve actionable results. There are many relevant methods, such as linear and quadratic programming, minimum squares solvers, and differential equation solvers (PDE, ODE). One individual method may be more convenient than another bank on the description of the enhanced function (linear, quadratic, or discontinuous) and restraints on the variables (linear or not). The objective is to produce results that start to valuable business choices. If the hospital staff knows a positive surgical form has high readmissions, they may advance the process to help scale down readmissions, such as granting for an extra day of post-operational care.

F. Step Six: Verify Results
After you applied your business resolution, allow time to production results. It is important to carefully approve the results facing the initial business objective. Returning to our healthcare illustration, the hospital’s business equitable is compressing re-missions. Analysts should analyse data to see if current rates have refused in a definite way.

G. Selecting the Right Tools
You may find your toolkit carried with a few appreciative software products to support the data investigation process, amid them analytic software that supplies mathematical and analytical algorithms. There are a few extensive benchmark to see such as scalability, reliability, performance, data source conformability, and comfort of distribution. When selecting a data analysis tool, it is important to contemplate these questions:

1. Is the tool memory-delimited?
2. Observe that decisive software should inform users of data failures. What if user input data is not applicable?
3. Check out the size of the problem; does it have a descriptive message to let the user know what is happening or would it bend the application?
4. Recognize supported data type, format, and environments. This add relational databases, structure and unstructured data, data connection backing, and
language backing. Does the tool support cascade data? Can the analytic be recycled inside the database?

5. In terms of achievement and technology, what is the development as well as the objective deployment environment? Will the analytics be thread secure? Does it support MapReduce (which will be essential for Hadoop)? Is the analytic software advanced for a deployment platform? Does it take dominance of multicore servers and can your computation be complemented?

6. What does the set up solution look like? Does it use industry standard original language to clarify embedding in your web, Linux or Windows application and deployment? Has it been tested over all platforms? If not, the computational results can be kind of different and origin differences in analytical results. Does it require any framework to support the formation? If so, what are the additional hardware, software, and maintaining costs?

Repetition is an attribute of the data process, not a component of the model. You can use anticipating analytics to go above hardly improving the ability of your current processes; you can create new convenience or products based on the vision you accumulated from the data. While this process seems difficult, there are matured, commercially-available tools that have been approved, tried, and in management, like Rogue Wave Software’s IMSL Numerical Libraries, to help companies appliance all six steps in this process. The IMSL Libraries afford refined analytics in high-performance, mission-critical operations. With IMSL, companies and organizations decrease development time, get a lower total cost of ownership, and better quality and conceivable.

![Figure 4: An applied conceptual architecture of big data analytics](image)

**References**


