Volume 4 Issue I, January 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Analysis and Prognosis of Cancer with Big Data Analytics

Manju.K.K[1], Mrs.Srinitya.G[2]
[1]PG Scholar, [2]Assistant Professor , Department of Information Technology,
Bannari Amman Institute of Technology, Sathyamangalam.

Abstract— Data Standards in cancer research have been evolving considerably. Technical Standards and data administration are the requirements of Diagnosis and Prognosis with enormous challenges. Cancer, a deadly disease can be analyzed with the innumerable genomic data and ethical analysis with big data functionalities as its crux. Normally the exome data for analysis is generated by NGS technologies. A single Cancer patient's exome data ranges from 10 Gigabytes to 15 Gigabytes. This largest range of Omics data can only be analyzed by Big Data computational models.
Keywords— Diagnosis, Prognosis, NGS, Exome data, Omics data.

I. INTRODUCTION

Cancer is still a major challenge in research and medical world despite many advances in technology. Analysis of cancer requires Petabytes of data. This requires various clinical and analytical databases with high dimensional scaling and interpretation. These databases are libraries of information that are collected from various scientific experiments, high throughput computational analysis and publication literature.Thereby, Biological and clinical researchers now face a increasingly large complex data sets. Although a standard genomic microarray may profile a genome for hundreds to thousands of features per sample. Current Next-Generation sequences provide over 100GB of raw sequence reads per genome. These data couples with clinical and phenotypic attributes have the potential to significantly expand our understanding of disease.However, this also provides Non-trivial issues in data storage and analysis.Thus, the requirement of Big Data and its computational technologies is the most required feature to scale these high dimensional datas.The analysis methods being incorporated with Big Data yields a result with the analysis and prediction over the survival rate of the patient.

The analysis will look for the genomics and clinical data from the patient as the input. The clinical information contains the details like age, tumor, node, metastasis and prescribed drug etc and genomics data contains DNA sequencing or Gene expression data. Our task is to identify the state of the disease and the survival potential of the patient. This also includes suggestions for line of treatment with possible drug, along with the side-effect and the toxicity of the prescribed drugs and the individual person specific correct medication by taking care of past medical history intelligently.

II. BIG DATA ANALYTICS IN HEALTH CARE

A Health data volume is expected to grow dramatically in the years ahead [1]. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in to-day's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits.

What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as "large volumes of high velocity, complex, and variable data that re-quire advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information". Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity [2]. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalysed) patient-related health and medical data to reach a deeper understanding of outcomes[3], which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

A. Research And Development

Comparative effectiveness research to determine more clinically relevant and cost-effective ways to diagnose and treat patients.
predictive modeling to lower attrition and produce a leaner, faster, more targeted R & D pipeline in drugs and devices;

statistical tools and algorithms to improve clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market; and Analysing clinical trials and patient records to identify follow-on indications and discover adverse effects before products reach the market.

B. Public Health

Analysing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response;

Faster development of more accurately targeted vaccines, e.g., choosing the annual influenza strains; and,

Turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of populations [4].

In addition, [5] big data analytics in healthcare can contribute to

1) Evidence-Based Medicine: Combine and analyse a variety of structured and unstructured data-EMRs, financial and operational data, clinical data, and genomic data to match treatments with outcomes, predict patients at risk for disease or readmission and provide more efficient care;

2) Genomic Analytics: Execute gene sequencing more efficiently and cost effectively and make genomic analysis a part of the regular medical care decision process and the growing patient medical record [6].

3) Pre-adjudication fraud analysis: Rapidly analyse large numbers of claim requests to reduce fraud, waste and abuse;

4) Device/Remote Monitoring: Capture and analyse in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction.

5) Patient Profile Analytics: Apply advanced analytics to patient profiles (e.g., segmentation and predictive modeling) to identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who would benefit from preventive care[5] .

III. TECHNOLOGY

A. Supervised Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables. Principal components analysis is commonly used as one step in a series of analyses. You can use principal components analysis to reduce the number of variables and avoid multicollinearity, or when you have too many predictors relative to the number of observations.

Supervised principal component analysis (supervised PCA), a generalization of PCA [7] that is uniquely effective for regression and classification problems with high-dimensional input data. It works by estimating a sequence of principal components that have maximal dependence on the response variable. The proposed supervised PCA is solvable in closed-form, and has a dual formulation that significantly reduces the computational complexity of problems in which the number of predictors greatly exceeds the number of observations (such as DNA microarray experiments). Furthermore, we show how the algorithm can be kernelized [8], which makes it applicable to non-linear dimensionality reduction tasks. Experimental results on various visualization, classification and regression problems show significant improvement over other supervised approaches both in accuracy and computational efficiency.

B. Cox Model

A Cox model is a statistical technique for exploring the relationship between the survival of a patient and several explanatory variables. Survival analysis is concerned with studying the time between entry to a study and a subsequent event (such as death). A Cox model provides an estimate of the treatment effect on survival after adjustment for other explanatory variables. In addition, it allows us to estimate the hazard (or risk) of death for an individual, given their prognostic variables. A Cox model must be fitted using an appropriate computer program (such as SAS, STATA or SPSS). The final model from a Cox regression analysis[9] will

yield an equation for the hazard as a function of several explanatory variables. Interpreting the Cox model involves examining the coefficients for each explanatory variable. A positive regression coefficient for an explanatory variable means that the hazard is higher and thus the prognosis worse. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable. The Cox model is a well-recognised statistical technique for analysing survival data. When it is used to analyse the survival of patients in a clinical trial, the model allows us to isolate the effects of treatment from the effects of other variables. The model can also be used, *a priori*, if it is known that there are other variables besides treatment that influence patient survival and these variables cannot be easily controlled in a clinical trial. Using the model may improve the estimate of treatment effect by narrowing the confidence interval. Survival times now often refer to the development of a particular symptom or to relapse after remission of a disease, as well as to the time to death. Such survival times are termed censored [10], to indicate that the period of observation was cut off before the event of interest occurred. From a set of observed survival times (including censored times) in a sample of individuals, we can estimate the proportion of the population of such people who would survive a given length of time under the same circumstances. This method is called the product limit or Kaplan–Meier method.

C. Kaplan–Meier Estimate Of The Survivor Function

To determine the Kaplan–Meier estimate of the survivor function, a series of time intervals is formed. Each of these intervals is constructed to be such that one observed death is contained in the interval, and the time of this death is taken to occur at the start of the interval. Some survival times are censored (that is, the patient did not die during the follow-up period). Sometimes the censored survival times occur at the same time as deaths. The censored survival time is then taken to occur immediately after the death time when calculating the survivor function. A plot of the Kaplan–Meier estimate of the survivor function is a step function, in which the estimated survival probabilities are constant between adjacent death times and only decrease at each death. An important part of survival analysis is to produce a plot of the survival curves for each group of interest. However, the comparison of the survival curves of two groups should be based on a formal non-parametric statistical test called the **logrank** test, and not upon visual impressions. The logrank test cannot be used to explore the effects of several variables, such as age and disease duration, known to affect survival. Adjustment for variables that are known to affect survival may improve the precision with which we can estimate the treatment effect. The regression method introduced by Cox is used to investigate several variables at a time. It is also known as proportional hazards regression analysis[9]. Briefly, the procedure models or regresses the survival times (or more specifically, the so-called hazard function) on the explanatory variable.

IV. WORKING PRINCIPLE

A. Architectural Framework

The conceptual framework for a big data analytics project in healthcare is similar to that of a traditional health informatics or analytics project. The key difference lies in how processing is executed. In a regular health analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition large, processing is broken down and executed across multiple nodes. The concept of distributed pro-cessing has existed for decades. What is relatively new is its use in analysing very large data sets as healthcare providers start to tap into their large data repositories to gain insight for making better-informed health-related decisions. Furthermore, open source platforms such as Hadoop/Map Reduce, available on the cloud, have encouraged the application of big data analytics in healthcare. While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional health analytics tools have become very user friendly and transparent. Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills. They have emerged in an ad hoc fashion mostly as open-source development tools and platforms, and therefore they lack the support and user-friendliness that vendor-driven proprietary tools possess. For the purpose of big data analytics, this data has to be pooled. In the second component the data is in a ‘raw’ state and needs to be processed or transformed, at which point several options are available. A service-oriented architectural approach combined with web services (middleware) is one possibility. The data stays raw and services are used to call, retrieve and process the data. Another approach is data warehousing wherein data from various sources is aggregated and made ready for processing, although the data is not available in real-time. Via the steps of extract, transform, and load (ETL), data from diverse sources is cleansed and readied. Depending on whether the data is structured or unstructured, several data formats can be input to the big data analytics platform.

In this next component in the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection and analytics models. Finally, on the far right, the four typical applications of big data analytics in healthcare

are shown. These include queries, reports, OLAP, and data mining. Visualization is an overarching theme across the four applications. Drawing from such fields as statistics, computer science, applied mathematics and economics, a wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyse, and visualize big data in healthcare.

B. Working

Omic data is used to create a classification model. When a new patient's data is received the Classifier uses classification model and patient data to classify whether patient has cancer or not. If a person is found infected, personalized drug will be suggested by the inference engine by accessing knowledge-base and clinical information.

Now applying the proposed method on dataset s1. There are 155 patients (columns) and 19794 dimensions (rows) in this data. Thus obtained the following results for dataset s1,

Applying k-means algorithm iteratively (for different values of k) until some threshold number of patients per clusters are found. For dataset s1, we defined threshold=8 and the total number of clusters calculated for this threshold are 4. For clusters 1, 2, 3 and 4 total number of patients identified are 25, 59, 8 and 63 respectively.

For all the 4 clusters identified using k-means algorithm ,apply SPCA[11] (Supervised Principal Component Analysis) for all 19794 dimensions repeatedly and identified the top genes that are responsible for each type of clusters. For identifying the top genes set the threshold value=1.99. With this specified threshold for the clusters 1, 2, 3 and 4 the number of top genes identified are 76, 33, 54 and 53 respectively.

Now, creating the graph database using neo4J where cancer is the root node and it leads to the breast cancer node (more specification) having genes as edge labels. From here breast cancer will be further divided into 4 types according to the 4 different clusters. And out of top genes that are identified for each clusters taking top 30 genes and these will be the label for 30 edges from breast cancer to this particular type of cancer.

After looking into the various open biological databases that have been mentioned previously we have to identify the drugs suitable for the particular type of cancer.

Finally a confined model named "Expert System" has been developed. As soon as a new patient's data arrives, it checks whether the person is suffering from cancer or not. If the person has cancer the system will suggest the suitable drug for it. If the person is already entering the prescribed drug then system can check whether the prescribed drug is toxic or not.

V. CONCLUSIONS

Big data analytics thus transformed the way of solutions in healthcare with its efficient and dynamic analysis. The supervised principal component analysis make the diagnosis well run and Cox regression model explores prognosis with its high interpretation over huge number of data.This is made more beneficial with its integration into Hadoop node. Thus enormous data can be scaled with the help of Hadoop environment, herby dynamic analysis is made possible.

REFERENCES

[1] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-report-download-today/.
[2] Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360; 2012. http://www.west-info.eu/files/big-data-in-healthcare.pdf.
[3] Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.
[4] Manyika J, Chui M, Brown B, Buhin J, Dobbs R, Roxburgh C, Byers AH: Big Data: The Next Frontier for Innovation, Competition, and Productivity. USA: McKinsey Global Institute; 2011
[5] IBM: IBM big data platform for healthcare." Solutions Brief; 2012. http://public. dhe.ibm.com/common/ssi/ecm/en/ims14398usen/IMS14398USEN.PDF.
[6] IBM: Large Gene interaction Analytics at University at Buffalo, SUNY; 2012. http://public.dhe.ibm.com/common/ssi/ecm/en/imc14675usen/IMC14675USEN.PDF.
[7] E. Bair and R. Tibshirani, "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data," PLoS Biol, p. 2(4): e108. doi:10.1371/journal.pbio.0020108, April 13, 2004.
[8] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. Technical report, 2004.
[9] Cox DR. Regression models and life tables. J Roy Statist Soc B 1972; 34: 187–220.
[10] Feldman B, Martin EM, Skotnes T: "Big Data in Healthcare Hype and Hope." October 2012. Dr. Bonnie 360; 2012.
[11] E. Bair and R. Tibshirani, "Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data," PLoS Biol, p. 2(4): e108. doi:10.1371/journal.pbio.0020108, April 13, 2004M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.

