A Survey of Big Data Analytics in Diseasomics Data For Precision Medicine

[1]T.Nagalakshmi, [2]M.Govindarajan

[1]Research Scholar, [2]Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Tamilnadu, India, [1]lakshusenthil@gmail.com, [2]govind_aucse@yahoo.com

Abstract: This paper gives the survey of the application of big data analytics in medicine and healthcare. Nowadays huge amounts of structured, unstructured and semi-structured data have been generated by many departments in worldwide which are collectively called as big data. Big data analytics in medicine and healthcare gives the analysis by integrating the heterogeneous data such as –omics data (datagenomics, epigenomics, transcriptomics, proteomics, metabolomics, diseasomics), electronic health records (EHR) data and biomedical data. A variety of big data analytics tools and techniques have been used for the analysis purpose and for giving precision medicine to patients. This survey gives the comparison of Classification and Clustering methods applied on the Cancer genomic data.

Keywords: Big data analytics, Precision Medicine, Bio informatics, -omics data, Biomarkers.

I. INTRODUCTION

1.1 BIG DATA and BIG DATA ANALYTICS

In recent years huge amounts of data is generated by different applications, wearable devices, geographical works, health industry and many more. Every industry needs analysis of their data. This massive amount of data creates many opportunities and gives challenges to the researchers. The big data is characterized by the following characteristics [1] : Value, Volume, Veracity, Variety, Variability. Figure 1 show the characteristics i.e. V's of Big Data.

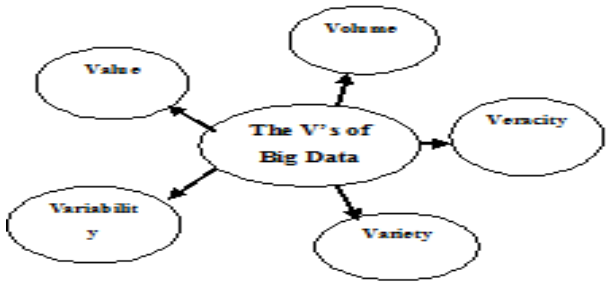

Figure 1: ***V's of Big Data***

The *volume* refers to the amount of data which is expected to be around terabytes, petabytes of data. *Veracity* refers to the relevance, quality, uncertainty and predictive value. *Variability* refers the consistency of data over a period of time. The collected data may be structured, semi-structured or unstructured which is referred as *variety*. The collected data is going to be analyzed which is referred as *value* of data.

The techniques and methods used for the traditional data are not adequate to handle the big data. The characteristics of big data impose challenges for the analysis of big data [2]. The volume of data generation and transmission of data are growing rapidly in recent days. The heterogeneous nature of big data imposes more challenges when they are used for analysis purpose. The traditional data analytics techniques are not sufficient for big data.

1.2 IMPACT OF BIG DATA ANALYTICS IN BIOINFORMATICS

Big data analytics plays an important role in Bioinformatics industry. The heterogeneous nature of data in healthcare sector imposes challenges in analyzing the data to give precision medicine. Big data analytics is used nowadays to predict the outcome of data from physicians, health status of patient, the outcome of an operation.

The following pathways will be the focus of the big data analytics industry [3]:

- Right living
- Right care
- Right provider
- Right value
- Right Provider

1.3 RESEARCH AREAS IN BIOINFORMATICS

The healthcare data include Electronic Health Records (EHR) data, machine generated/sensor data, patient registries, genetic databases, and public records. Public records are the major sources of big-data in the healthcare industry. It requires efficient data analytics to solve their associated healthcare problems.

[1] listed different –omics data and the importance of the study about each data. In Table 1, the various –omics data and the aim of the study about each data is listed.

–Omic and EHR big data analytics is challenging due to[4]*:*

 DOI : 10.35291/2454-9150.2019.0039

- Diverse data collection frequency
- Inherent data quality issues
- High dimensionality
- Heterogeneous data types

Table 1 The main aim of various classes of –omics data

-omics data	Aim of the data
Geneomics	Study of genomes of organisms, incorporates elements from genetics.
Proteomics	Study of proteins.
Metabolomics	Study of small molecules i.e metabolites
Diseasomics	Study of all diseases and disorders of an organism.
Metalobomics	Study of the whole set of the metabolites (small-molecule compounds) within a cell, a tissue or an organ.
Transcriptomics	Study of the expression level of all RNAs in particular cell, or cell population

Some –omic data, a genome is invariant over a long period of time, and often only needs a one-time data acquisition. But the other types of –omic data vary with environment, tissue types, and time that require multi time-point acquisition.

A big challenge in –omic and EHR data is the high-dimensional nature of the data. According to the heterogeneity of the data sources, the noise of the experimental – omics data and the variety of the experimental techniques, environmental conditions and biological nature should be considered, before integrating the heterogeneous data and before applying data mining methods.

Even though the various big data analytic tools and techniques used in recent years, the health care industry still suffers the big challenge of giving precision medicine to patients. Precision medicine requires data utility ranging from collection and management to analytics of big data.

Why the precision medicine to each patient is not efficient in this sector? The correct reason is a portion of new biomarkers are patented and that too adopted in small portion / small percentage in standard clinical practices [5]. So, the healthcare industry and big data analytics field must provide the tools and techniques to give optimizing care for the specific needs of each patient.

The tools used so far the identification of biomarkers is listed in Table 2. For genomic data, GWAS (genome-wide association studies) uses different approaches (e.g., the chi-squared test or logistic regression) to find the degree of association between each variant and a targeted attribute, and then select most significant variants as biomarkers.

Table 2 Tools for –omic biomarker identification.

Tool	–Omic Data	–Omic Biomarker
SNPassoc	Genomic	Significant SNPs associated with attributes
CNVRuler		Significant CNVs associated with attributes
edgeR∗	Transcriptomic	Differentially expressed genes /transcripts
DetectTLC	Proteomic and metabolomic	Molecular patterns in mass spectrometry images

II. LITERATURE SURVEY

2.1 CANCER GENOME PROJECT

Cancer is considered to be as the most complex disease in recent years. More than 200 forms of cancer have been discovered and each type can be characterized by different molecular profiles.

Public cancer genome projects are:

- International Cancer Genome Consortium (ICGC)
- The Cancer Genome Atlas (TCGA)
- Asian Cancer Research Group (ACRG)

The Cancer Genome Atlas (TCGA) is a public funded project to create an "atlas" of cancer genomic profiles from cancer causing genomic alterations [6]. TCGA researchers so far discovered and analyzed large cohorts of over 30 human tumours through large-scale genome sequencing and integrated multi-dimensional analyses. The National Institute of Health (NIH) launched TCGA Pilot Project to create an "atlas" of cancer genomic profiles. Providing publicly available cancer genomic databases will permit the improvement of diagnostic methods, treatment standards, and finally cancer prevention.

TCGA has Research Network Center that includes several cooperating centers for processing the samples and managing all the obtained datasets.

TCGA project engaged by scientists and managers from NIH's NCI (National Cancer Institute) and NHGRI (National Human Genome Research Institute). They are concentrating on the projects like:

- Glioblastoma
- Breast Cancer
- Ovarian Cancer
- Lung Cancer
- Gastric adenocarcinoma.
- Pan-Cancer projects

 DOI : 10.35291/2454-9150.2019.0039

[7] discussed about the reasons for getting the risk of cancers. Living pattern, excess food & no weight control, no exercise & physical work and beverages of alcohol are the main causes for getting cancer. The hormone based analysis showed that the hormones & growth factors are responsible for the risk of Breast cancer.

Since hormone measurement is not feasible for post-menopausal women, they have chosen estradiol and testesterone as the good biomarker for the promising weight loss.

Usually 5 kg/m^2 increase in BMI increases the risk of breast cancer about 12% and weight reduction of 5% reduces the risk factor to 25% to 40%. Risk prediction is mainly based on the number of risk factors woman is carrying. Mammographic density helps in identification of risk factor when compared with the standalone models. According to mammographic density in breast, dense tissue is always white and fat tissue is radio-lucent & appears black.

[8] discussed about the research in classification of diseases and symptoms of breast cancer. They listed the following symptoms for a breast cancer:

- A lump in a breast.
- A rash around one of the nipples.
- A swelling (lump) in one of the armpits.
- An area of thickened tissue in a breast.
- The size or the shape of the breast changes.
- The nipple changes in appearance; it may become sunken or inverted.
- A pain in the armpits or breast that does not seem to be related to the woman's menstrual period.

2.2 DATA ANALYTICS TECHNIQUES FOR CANCER DATA

Data analytics has various techniques such as pre-processing, classification, clustering, prediction, association rules, and regression. Many researchers have attempted to apply machine learning algorithms for detecting the cancers in human beings.

There are 2 major classifications of cancer data; *malignant* which is dangerous & cancerous when a start growing inside, benign is less harmful as its cells don't multiply. So the early detection of anomalies in breast enables the doctor's in diagnosing the breast cancer easily which can save several lives. [9] used Decision Tree & Random Forest methods for the early detection of anomalies in breast. The effectiveness of the classification algorithms are measured using the evaluation metrics such as Accuracy, Specificity and Sensitivity.

[10] compared the classification algorithms namely Naive Bayes, Multilayer Perceptron, Radial basis function network, nearest neighbour, Conjunctive rule algorithms and have done early prediction of breast cancer by using these methods. The author concluded that Naïve Bayes algorithm gives better performance in early detection. Confusion matrix is used for his evaluation.

Data pre-processing is vital for multi-omics integrative analyses in terms of reducing unwanted biases and noises. [11] have developed new feature selection algorithm which considers the interdependencies among features. They developed a graph based solution for the interdependencies problem. They concluded that this new method gives the best feature selection because the existing feature selection methods failed to represent the original data since they assumes that the features are independent.

[12] gave the comparison of classification techniques like Decision tree, K-Nearest neighbor, SVM, Baysian network & Naïve bayes algorithms. They concluded that the Baysian Network gives best accuracy with less featured datasets and SVM gives best accuracy with more features datasets.

[13] enhanced the Euclidian distance formula to increase the cluster quality in K-Means clustering technique. This new algorithm is concentrating on the enhancement of clusters that is based on normalization. Two new features are added for the enhancement. The first point is to calculate normal distance metrics on the basis of normalization. In second point the functions will be clustered on the basis of majority voting.

[14] classified the cancer datasets using three categories of integrative clustering methods. They compared Direct Integrative Clustering, Clustering of Clusters (COC) and Regulatory Integrative Clustering techniques. Finally they concluded that integrative clustering methods should take more considerations on the increasing computational burden like memory requirement, parallel computing ability, time and cost in future works.

[15] has taken cancer datasets from TCGA public funded project. They have taken 11,188 patients human samples which represents 33 different cancer types. They performed integrative clustering method iCluster technique on the chosen data. They performed molecular clustering using data on chromosome-arm-level aneuploidy, DNA hypermethylation, mRNA,and miRNA expression levels and reverse-phase protein arrays, of which all, except for aneuploidy, revealed clustering primarily organized by histology, tissue type, or anatomic origin.

[16] has used unsupervised classification of proteins and metabolites using Deep Learning (DL) methods such as LSTM-VAE & DCEC (deep convolutional embedded clustering). They compared the DL method with the integrative analysis (K-Means, PAM & Hierarchical Clustering methods) methods. The authors have concluded that DL methods found meaningful Clusters in multi-omics data.

 DOI : 10.35291/2454-9150.2019.0039

2.3 CANCER BIOMARKERS AND PRECISION MEDICINE

There are multiple similar genes which could be gnomically or epigenetically changed since they depend on mutations, copy number variations, epigenetic modifications. Therefore, different individuals who share the same phenotype/diseases may have different causal genes. Thus, they may require different drug targets. We should give the 'right drug' to the 'right patient' at the 'right time' [17]. One of the goal of many ongoing precision medicine programs is to reach this goal using –omic data or in combination with environmental / lifestyle factors by identifying the biomarkers.

Multiple Survival Screening (MSS) and Significance Analysis of Prognostic Signatures (SAPS) methods have been developed for identifying cancer biomarkers. A portion of new biomarkers are patented and that too adopted in small portion / small percentage in standard clinical practices. There are more challenges in identifying biomarkers for each type of cancer.

III. PROPOSED METHOD

Data analytics has various techniques such as pre-processing, classification, clustering, prediction, association rules, and regression. Until now, all these techniques were applied on the medical data and analyzed [18][19]. Preprocessing work is needed for the EHR medical data. Even though the various big data analytic tools and techniques used in recent years, the health care industry still suffers the big challenge of giving precision medicine to patients. In this proposed method, there are 3 commitments are going to be done on the breast cancer data. First commitment towards analyzing the cancer genome data is the post-processing works on the –omic data, especially in cancer genome data and will be analyzed. There are 2 major classes of cancer data: malignant and benign. Second commitment for the –omic big data is that the classification techniques has to be applied after clustering of big data i.e. after forming good clusters. Third commitment is that all the identified biomarkers are still not practiced to give precision medicines to patients. Still the healthcare industries especially the cancer datasets require new biomarkers for the benefit of patient [20]. This website includes the types of cancer data, treatment for the cancer types and technologies used until now to give the precision medicine.

IV. CONCLUSION

The healthcare data include Electronic Health Records (EHR) data, machine generated/sensor data, patient registries, genetic databases, and public records. Public records are the major sources of big-data in the healthcare industry. There are 2 major classifications of cancer data; *malignant* which is dangerous & cancerous when a start growing inside, *benign* is less harmful as its cells don't multiply. Usually 5 kg/m^2 increase in BMI increases the risk of breast cancer about 12% and weight reduction of 5% reduces the risk factor to 25% to 40%. So, the healthcare data and big data analytics are the big research areas. By delivering the most suitable and effective treatment to each patient based on their precise health information, the healthcare system can achieve better efficiency and quality.

REFERENCES

[1] Ristevski, B. & Chen, M. "*Big Data Analytics in Medicine and Healthcare*" in Journal of Integrative Bioinformatics, 15(3), 2018.

[2] Kashyap, H., Ahmed, H.A., Hoque, N. et al. "*Big data analytics in bioinformatics: architectures, techniques, tools and issues*" in Network Modeling Analysis in Health Informatics and Bioinformatics, 2016, 5: 28.

[3] S. Kumar and M. Singh, "*Big data analytics for healthcare industry: impact, applications, and tools*," in Big Data Mining and Analytics, 2(1), 2019, pp. 48-57.

[4] P. Wu, C. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang, *"–Omic and Electronic Health Record Big Data Analytics for Precision Medicine*," in IEEE Transactions on Biomedical Engineering, 64(2), 2017, pp. 263-273.

[5] A. S. Panayides, M. Pattichis, S. Leandrou, C. Pitris, A. Constantinidou and C. S. Pattichis, "*Radiogenomics for Precision Medicine With A Big Data Analytics Perspective*," in IEEE Journal of Biomedical and Health Informatics, 2018.

[6] Tomczak K, Czerwińska P, Wiznerowicz M. The *Cancer Genome Atlas (TCGA): an immeasurable source of knowledge*. Contemp Oncol (Pozn). 2015; 19(1A):A68–A77.

[7] Shalaka Ghaisas and Shine Devarajan, "*A Review of the Predictive Aspects of Breast Cancer among Women*", Intl. J. Bioinformatics and Biological Sci.: 6(1), 2018, p. 25-33.

[8] Disha Patel, Bhavesh Tanwala, Pranay Patel, "*Breast Cancer Using Data Mining Techniques*", International Journal of Computer Sciences and Engineering, 6(7), 2018, pp.1531-1536.

[9] N. Sridevi, S. Anitha , "*Prediction of Breast Cancer using Decision tree and Random Forest Algorithm*", International Journal of Computer Sciences and Engineering, 6(2), 2018, pp.226-229.

[10] Tamilvanan, B. (2018). "*An Efficient Classifications Model For Breast Cancer Prediction Based On Dimensionality Reduction Techniques*", International Journal of Advanced Research in Computer Science, 9, 2018, pp. 448-455.

 DOI : 10.35291/2454-9150.2019.0039

[11] T. B. Mudiyanselage and Y. Zhang, "*Feature selection with graph mining technology*," in Big Data Mining and Analytics, 2(2), 2019, pp. 73-82.

[12] Ajay Kumar, R. Sushil, A. K. Tiwari, "*Comparative Study of Classification Techniques for Breast Cancer Diagnosis*", International Journal of Computer Sciences and Engineering, 7(1), 2019, pp.234-240.

[13] A. Bansal "*Improved K-mean clustering algorithm for prediction analysis using classification technique in data mining*" in Int. J. Comput. Appl., 157, 2017, pp. 35-40

[14] Wang, D. & Gu, J. Quant Biol (2016) "*Integrative clustering methods of multi-omics data for molecule-based cancer classifications*" in Quantitative Biology 4: 58.

[15] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf *et al*., "*Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer*", in Cell, 173(2), 2018, Pages 291-304.

[16] Neo Christopher Chung, Bilal Mirzaa, Howard Choi et al. "*Unsupervised classification of multi-omics data during cardiac remodeling using deep learning*" in Methods, 2019.

[17] Wang E, Cho WCS, Wong SCC, Liu S. "*Disease Biomarkers for Precision Medicine: Challenges and Future Opportunities*" in Genomics Proteomics Bioinformatics,15(2), 2017, pp.57–58.

[18] Nagaraj, K., Sharvani, G.S.and Sridhar, A. '*Emerging trend of big data analytics in bioinformatics: a literature review'*, Int. J. Bioinformatics Research and Application, 14(1/2), 2018, pp.144–205.

[19] S. Bahri, N. Zoghlami, M. Abed and J. M. R. S. Tavares, "*BIG DATA for Healthcare: A Survey*," in IEEE Access, vol. 7, 2019, pp. 7397-7408.

[20] Accessed: April, 2019. [Online]. Available: https://www.cancer.gov/about-cancer/treatment/types/precision-medicine

 DOI : 10.35291/2454-9150.2019.0039