INCEDO
WHERE INNOVATION PROPELS

CANCER PREDICTION and Prognosis - The Machine Learning Way!

Cancer is usually characterised as a heterogeneous disease with different types and subtypes. Cancer is not a single disease, but rather many related diseases that involve uncontrolled cellular growth and reproduction. It is one of the leading causes of death in the developed world and second in the developing world, almost a million people around the globe are losing the fight against cancer every year. The early diagnosis and prognosis of a cancer type have become inexorable in cancer research, as it can facilitate the quality of life of patients. Classifying cancer patients into high or low risk groups is the most important task that has led many research teams, be it from the biomedical or the bioinformatics field, to study the application of Machine Learning (ML) methods and to improvise on it. ML techniques have been exploited as an aim to model or to simulate the progression and treatment of cancerous conditions. In addition, ML tools have the ability to detect key features from complex datasets, thus assert their importance. ML has broad range of these techniques, including Artificial Neural Networks, Bayesian Networks, Support Vector Machines and Decision Trees. These methods have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making with less time and investment.

> “ML techniques have been exploited as an aim to model or to simulate the progression and treatment of cancerous conditions ”

ML is the subfield of computer science that studies programs that generalize past experience. This project looks at classification, where an algorithm tries to predict the label for a sample. A sample is a single set of feature data (here, gene expression levels for a cancer patient) plus a label, which is what category (for example, basal or luminal) the sample falls in. The machine learning algorithm takes many of these samples, called the training set, and builds an internal model. Using this model, it can, then, predict the labels of other samples, called the testing set.

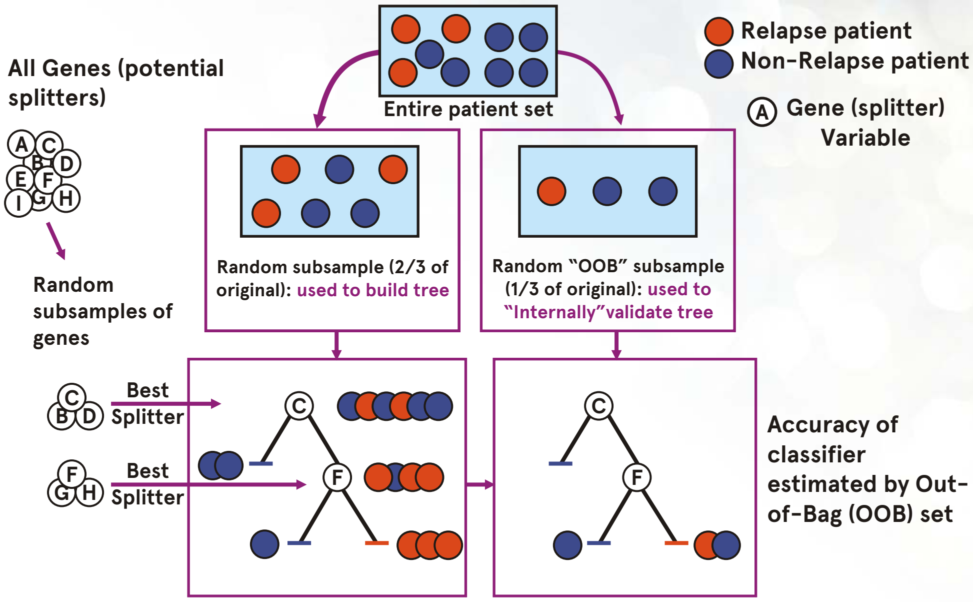
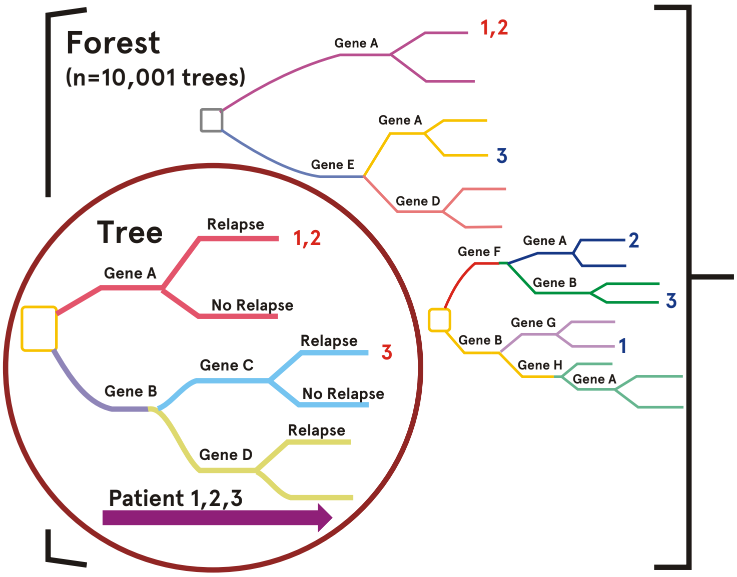

Relapse patient
Non-Relapse patient
A Gene (splitter) Variable
All Genes (potential splitters)
A C B D E F I G H
Random subsamples of genes
Entire patient set
Random subsample (2/3 of original): used to build tree
Random "OOB" subsample (1/3 of original): used to "Internally"validate tree
C B D Best Splitter
F G H Best Splitter
C
F
C
F
Accuracy of classifier estimated by Out-of-Bag (OOB) set

Figure 1: A Random Forest is built one tree at a time.



The more often a gene is chosen as a splitter variable, the higher its "Variable Importance"- This can be used to prioritize which genes to select for an assay with limited gene measurements

Gene	Var. Imp.
Gene A	0.67
Gene B	0.20
Gene D	0.13
...	...

Figure 2: Variable importance is a feature of random forests

Various ML techniques and feature selection algorithms have been widely applied to disease prognosis and predictio. ML techniques can predict cancer susceptibility, recurrence and survival. With the advent of genomic, proteomic and imaging technologies huge molecular information is obtained. Molecular biomarkers, expression of certain genes are a crucial indicator in cancer prediction. Huge amount of data generated from high throughput sequencing technologies is made available for exploration by the scientific communities. ML methods can be successfully implemented to discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type. This can be used to effectively predict future outcomes of a cancer type. There are many ML tools which are capable of biomarker prediction and disease classification. Further, ML can be used for revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection.

Machine Learning For Cancer Classification - Part 2 - Building A Random Forest Classifier by Nicholas Spies: https://www.biostars.org/p/86981/

The development of a community resource project, namely The Cancer Genome Atlas Research Network (TCGA), has further provided large scale genomic data about specific tumor types in public domain [11]. TCGA provides with the ability to better understand the molecular basis of cancer through the application of high-throughput genome technologies.

In 2012, a publication proposed the use of ML techniques for cancer recurrence prediction. The recurrence prediction of oral squamous cell carcinoma has exploited heterogeneous sources of data (clinical, imaging and genomic) in order to predict a possible relapse of OSCC and a subsequent recurrence. Park K.et.al developed a predictive model for the evaluation of survival in women who have been diagnosed with breast cancer. They compared three classification models namely SVM, ANN and SSL based on the SEER cancer database.

ML Applications - Breast Cancer Samples

Incedo research team has evaluated a large scale population based next generation sequencing data of breast cancer provided in TCGA repository. We have screened RNAseq (~1200) samples from TCGA. These samples were segregated with respective disease and control. An end-to-end analysis was carried out which includes quality control, gene quantification, normalization and differential gene expression. Outcomes were further tested on trained vector machine to evaluate risk association of genes in cancer disease prognosis and to identify probable biomarkers. We have selected the RNAseq data which is less noisy and can improve the predictive performance of classification algorithms. Further, we have identified few novel transcripts and genes which may be considered as biomarker in our breast cancer samples. Our analysis revealed that Support Vector Machine and Random Forest showed similar predictive performances and both methods correctly classified samples with 83.3% classification accuracy. These decision support tool developed for RNA-Sequencing datasets will assist researchers in their decisions for diagnostic biomarker discovery and classification problem.

For detailed case study, please visit us at: www.incedoinc.com

Conclusion

The concepts of ML and their application in cancer prediction/prognosis are profound. However, further improvement depends on data diversity and enrichment for training the machines. Most of the studies focus on the development of predictive models using supervised ML methods and classification algorithms. Integrating ML into scientific domain makes a great difference in critical decision making. It is evident that the integration of multidimensional heterogeneous data with application of different ML technique for feature selection and classification can be a great asset to human understanding in cancer domain.

Reference

1. Kourou, K., et al., Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J, 2015. 13: p. 8-17
2. Hanahan, D. and R.A. Weinberg, Hallmarks of cancer: the next generation. Cell, 2011. 144(5): p. 646-74.
3. Cruz, J.A. and D.S. Wishart, Applications of machine learning in cancer prediction and prognosis. Cancer Inform, 2006. 2: p. 59-77.
4. Cicchetti, D.V., Neural networks and diagnosis in the clinical laboratory: state of the art. Clin Chem, 1992. 38(1): p. 9-10.
5. Cochran, A.J., Prediction of outcome for patients with cutaneous melanoma. Pigment Cell Res, 1997. 10(3): p. 162-7.
6. Exarchos, K.P., Y. Goletsis, and D.I. Fotiadis, Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed, 2012. 16(6): p. 1127-34.
7. Kononenko, I., Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med, 2001. 23(1): p. 89-109.
8. Ren, X., et al., ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. Nucleic Acids Res, 2013. 41(4): p. e53.
9. Ren, X., et al., iPcc: a novel feature extraction method for accurate disease class discovery and prediction. Nucleic Acids Res, 2013. 41(14): p. e143.
10. Wang, Y., et al., Revealing metabolite biomarkers for acupuncture treatment by linear programming based feature selection. BMC Syst Biol, 2012. 6 Suppl 1: p. S15.
11. Zhang, K. and H. Wang, [Cancer Genome Atlas Pan-cancer Analysis Project]. Zhongguo Fei Ai Za Zhi, 2015. 18(4): p. 219-23.
12. http://seer.cancer.gov/about/
13. Kourou, K., et al., Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J, 2015. 13: p. 8-17

GOPAL JOSHI
Scientist - Bioinformatics | Pharma and Life Sciences Practice
Email: gopal.joshi@incedoinc.com
https://in.linkedin.com/in/gopsa

ATEEQ KHALIQ
Scientist - Bioinformatics | Pharma and Life Sciences Practice
Email: ateeq.khaliq@incedoinc.com
https://in.linkedin.com/in/ateeqkhaliq

SANDEEP PUNJANI
Director – Life Sciences Practice
Mob: +1 (201) 317-8291
Email: sandeep.punjani@incedoinc.com
https://www.linkedin.com/in/punjani

About us

Incedo Inc (formerly a part of $4Bn Indiabulls Group) is a technology solutions and servicing organization headquartered in the Bay Area, USA with workforce across North America, South Africa and India (Gurgaon, Bangalore). We specialize in Data & Analytics and Product Engineering Services, with deep expertise in Financial Services, Life Science and Communication Engineering. Our key focus is on Emerging Technologies and Innovation. Our end-to-end capabilities span across Application Services, Infrastructure and Operations. What really differentiates us is:

- Strong engineering talent
- Focus and passion for innovation
- Flat organization structure – responsive engagement models
- Agile and flexible delivery and commercial models
- Focus on long term partnership with clients

incedo
WHERE INNOVATION PROPELS

USA: | 2350 Mission College Boulevard, Suite 246 Santa Clara, California - 95054 | Tel: +1408 531 6040
INDIA: | 248, Udyog Vihar Phase-IV, Gurgaon - 122 015 | Tel: +91 124 4345900/01/02
www.incedoinc.com