International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

A Comprehensive Study of Machine Learning Models in Radiogenomics

Eash Sharma[1], Ashwin Garg[2]

[1, 2]Biochemical Engineering and Biotechnology, IIT Delhi

Abstract: ***The ever increasing medical data has led to an increasing interest and demand for a personalized treatment setup in which each individual has its own personalized treatment plan. Specifically talking, Radiation Oncology has generated a lot of input as well as output data through which it has been able to capture the interest of the Machine Learning Methodologies. Going further, Radiogenomics , in particular, the study of genetic variation associated to radiation has been seen as a potentiate user of a lot of Machine Learning approaches. Currently, uniform doses specific to the tumor are being used. The contribution of genetics to radiations far exceeds the current understanding of risk variants. In this paper, we study the applications of Machine Learning in the Radiogenomics field which have been compared and contrasted to overcome the shortcomings of the current situation.***

Keywords: Radiogenomics, Machine Learning, Personalized Treatment, Radiation Oncology

1. Introduction

Current dose practices are suboptimal. First, the dose is constrained by the surrounding normal tissue. Second, there can be 2 similar dose distributions among patients with different toxicities resulting in suboptimal treatment which fails to give a good quality of life. Ml methods being used will be able to detect knowledge from deeper levels. The models introduced depict the importance of integration of ML with radiogenomics in the future.

2. Radiogenomics

2.1 Normal tissue toxicity

Therapeutic radiation has been seen to deliver an effective maximal dose while minimizing the normal tissue toxicity. Previously fatal cancers have become curable and patients have had to live with malignancies. Also, acute toxicity results in constrained dose escalation which in turn results in limited tumor control. Improvements in therapeutic ratio have been made due to technological advances like IMRT,3-D Planning using CT simulation. In recent years, through various studies based on patients, it has been seen that genomic factors could influence susceptibility for the development of radiation related toxicities. To identify these genomic factors, a series of gene studies were performed. The findings did not come to a conclusion however. The risk of SNPs has been a concern for these genome studies. To improve methods to detect new SNP markers for radiation toxicity, REQUITE, a project led by RGC members to collect biological data and genetic information for cancer patients has been devised. This provides us with a huge amount of data which needs to be worked on.

2.2 Fundamental Hypothesis of Radiogenomics

The three basic hypothesis by Andreassen show that there are epigenetic components of normal tissue radiosensitivity that are not captured by genetic sequences but are heritable.

2.3 Machine Learning

Encompassing Computer Science, Statistical inference and artificial intelligence, machine learning seeks to uncover patterns in date to make future predictions.

2.3.1 Statistical Inference vs ML

Machine learning has a considerable overlap with classical statistics. In ML, models are measures of predictive performances whereas statistics values model according to the goodness to fit. One key difference is the absence of formal hypothesis testing in ML.

2.3.2 Breiman's Lessons from ML

Breiman noted three lessons from ML which have relevance to contemporary issues of ML usage in medicine.

2.3.2.1 Rashomon Effect

This effect describes a multiplicity of models that have a very similar performance but very different compositions. This model effect is magnified by feature selection as the remaining variables must then implicitly carry the effect of the removed variables.

2.3.2.2 Occam Dilemma

This dilemma focuses on the choice between simplicity and accuracy. It has been noted that decision trees and logistic regression have been relatively simple but were outperformed by more complex classifiers like random forests.

2.3.2.3 The curse of dimensionality

This refers to the phenomenon where potential data space increases exponentially with increasing number of dimensions.

3. Methods

Nowadays, radiogenomics uses ML techniques in the top-down approach, where the outputs use complex statistical analysis for modelling. And not taking into account *a priori* knowledge of interactions of radiation with various biological systems. In the field, we usually prefer supervised

Licensed Under Creative Commons Attribution CC BY

learning, which means that the predictive hypothesis and model are generated from a labeled set of training examples. Although, in some cases unsupervised learning might also be preferred. Also, feature selection is of extreme importance that is determining which features are the most important.ML techniques consist both of regression and classification models. In our interest generally lies the classification models usually classifying the data into presence/absence of various disease and therapy features. Also, one of the keen interest lies in overcoming the Occam Dilemma and the use of techniques in such a way that allows ready interpretation. Keeping this in mind, our main focus of study in the section will be Random Forests, Support Vector Machines and Bayesian Networks.

Table 1: Three representative machine learning techniques

Method	Pre-process	Complexity control
Support Vector Machine(SVM)	• Encode features as binary	• Recursive feature elimination for linear SVM
	• Normalize to uniform distribution	• Soft margin width(C-parameter)
	• Imputation for balancing data	• Kernel hyperparameters
Bayesian networks	• Feature discretization	• Constraints to a graph search space based on prior knowledge
	• Variable selection to reduce graph search space	• Graph scoring functions that penalize completely
	• Imputation not necessary when using expectation maximization	
Random forest	• No discretization or normalization necessary.	• Number of features to sample at each node split
	• Imputation required	• Minimum number of samples in a terminal node

3.1 Random Forest

A regression and classification based approach based on a group of decision trees. Each tree is trained on bootstrapped training examples and a random subset of features is used at every node split. For example, when this method is applied to predicting disease using SNPs (Single Nucleotide Polymorphisms), each tree grows with a fixed set of rules to divide the training samples that are based on genotypes.

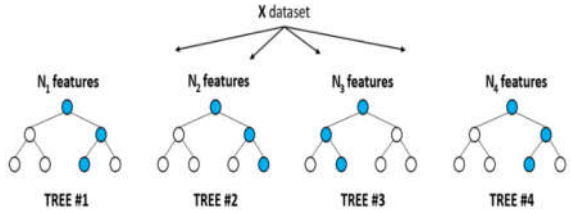

Figure 1: Random Forest

3.1.1 Robustness

For high dimensional data, there is always a risk of overfitting. The aggregation of the trees are based on low correlation which minimizes the risk by reducing the model variance. When training RF models, some parameters need to be optimized, which affect the model power. For example, the number of variables that are to be selected at the node splits. Many studies select default configurations as originally recommended by Breiman.

3.1.2 Ability to account for SNP SNP interactions

Epistasis, the non linear combination of SNPs that may correlate with a phenotype is very important for understanding complex diseases. RF indirectly accounts for node splits as one node split is conditional upon the previous node split. As a result, RF has been used as a screening step to identify much smaller number of SNPs that demonstrate SNPs.

3.2 Support Vector Machines

Used to classify patients into 2 separate classes based on their characteristics. The first step includes finding an efficient boundary between the 2 groups. A technique called Kernel trick is used to determine this boundary. SVMs maximize the distance between the 2 classes and allow a defined number of cases to be on the wrong side of the boundary. Due to this, SVMs are only minimally influenced by outliers that are difficult to separate.

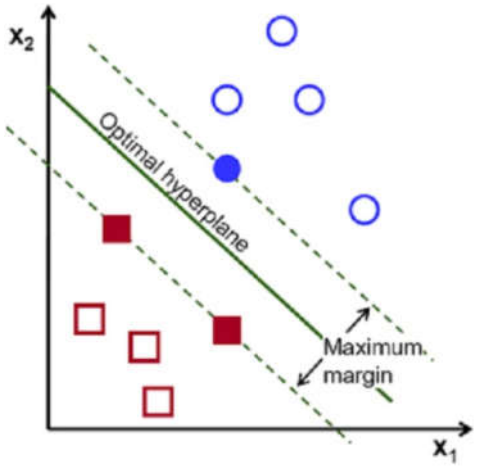

Figure 2: Support Vector Machine

3.2.1 Robustness

SVMs have a possible complex and unknown correlation structure by means of adaptable non-linear classification boundaries. SVMs can be used to tackle GWAS data in a combination of steps.A 2 step SVM procedure with SVMs first adopted for testing SNPs by taking the correlation structure into account and identifying a subset of relevant candidate SNPs. Subsequently, statistical hypothesis testing is performed with an adequate threshold correction.

3.2.2 Tuning Parameters

Some key issues in SVM modeling are tuning the parameters, identifying the separating hyperplane and the number of vectors that must be used. Also, some kernel specific tuning is required.

Licensed Under Creative Commons Attribution CC BY

3.3 Bayesian Networks

A graphical method that is used to model the joint probabilistic relationships among the set of random variables. Based on the analysis of the input data, BN assigns probability factors to the various results. An integral component of BN is DAG(directed acyclic graph).A DAG is made up of nodes and links between them. The probability of each random variable is conditioned upon its parent variable.

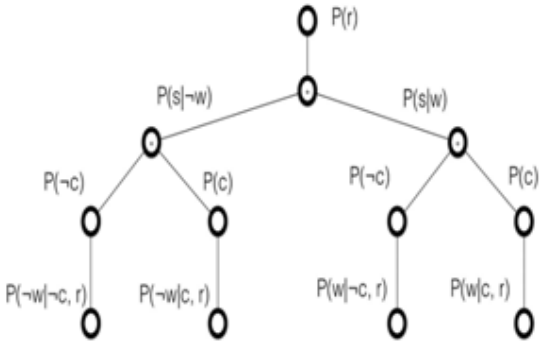

Figure 3: Bayesian Network

3.3.1 Robustness

As the number of DAGs grows exponentially with the number of features, it is not feasible to search for the highest scoring DAG over all possibilities. Some approaches to tackle the problem are reducing input dimension and a prior use of causality that considers the knowledge at hand to impose restrictions on the direction of links between nodes to reduce the search space.

3.3.2 Using the data and knowledge

A DAG can be built starting from previous knowledge or completely trained with the available data.The optimized DAG is one which maximizes a predefined scoring function over all possible DAG configurations.

4. Literature Survey

We have used these models to predict an optimal level of treatment and dosage with the integrated use of Machine Learning and Radiogenomics.. We chose the problem so as to predict an optimal level of dosage for patients in the future as the huge amount of past patient data can be dug out as a meaningful indicator for future endeavors. The presented algorithms can accommodate GWAS-level data. When we consider this emerging sequencing domain, new technical challenges are posed which further can be addressed by some new advances in the algorithm world.

5. Results and discussion

Machine learning promises a significant advance in radiogenomics knowledge. In this section, we will have a look at the general lessons learned as well as the potential barriers.

5.1 Lessons from statistics

ML models focusing alone on predictive performance and not admiring the importance of statistical theory would be a mistake. Through many iterations, statistics has learned that it is crucial to take into consideration multiple hypothesis testing to decrease type 1 error. Although, ML models are trained to be hypothesis-free, they often fall into a trap of cherry picking results that show a good performance, which might be spurious. The phenomenon is not rare in Oncology as a result of the desire to find an application for a therapeutic. A proposed solution is creating drug development portfolios to apply meta-analysis principles to drug trials instead of regarding them as individuals. A similar approach could be used in Radiogenomics to avoid bias.

5.2 Incorporating clinical variables

Most of the disease phenotypes are confounded by the environment. When genetic and environmental effects are combined, there is increased accuracy in heritability prediction. This suggests that models encompassing both genetic and clinical factors should offer a superior prediction. Current models do not incorporate genetic factors.

5.3 Replication and Regulatory Concerns

While applying these ML based algorithms, we also need to take into account the current regulatory environment.There is a controversy regarding whether and how the FDA should regulate laboratory-developed tests while still promoting innovation. One possible solution is pre-certifying laboratories instead of instead of individual LDTs.

6. Conclusion

Computational Oncology is a field based on rich multidisciplinary study and currently the focus on machine learning in the field is quickly moving towards medicine. However, the translational research efforts in field are difficult and require coordinated work from various stakeholders belonging to varied backgrounds. With the advent of the genomic era, the importance of machine learning in the field of Oncology will only increase further. Hence, insights from computational biologists, statistical geneticists, and ML researchers will play a crucial role in the field of computational oncology in the future.

References

[1] Hall EJ, Giaccia AJ. *Radiobiology for the Radiologist*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins (2012).
[2] Mould RF. Pierre curie, 1859–1906. *Curr Oncol* (2007) 14(2):74–82. doi:10.3747/co.2007.110
[3] Baumann M, Krause M, Overgaard J, Debus J, Bentzen SM, Daartz J, et al. Radiation oncology in the era of precision medicine. *Nat Rev Cancer*(2016) 16(4):234–49. doi:10.1038/nrc.2016.18
[4] Chun SG, Hu C, Choy H, Komaki RU, Timmerman RD, Schild SE, et al. Impact of intensity-modulated radiation therapy technique for locally advanced non-small-cell lung cancer: a secondary analysis of the NRG oncology RTOG 0617 randomized clinical trial. *J Clin*

Licensed Under Creative Commons Attribution CC BY

Oncol (2017) 35(1):56–62. doi:10.1200/JCO.2016.69.1378

[5] Folkert MR, Singer S, Brennan MF, Kuk D, Qin LX, Kobayashi WK, et al. Comparison of local recurrence with conventional and intensity-modulated radiation therapy for primary soft-tissue sarcomas of the extremity. *J Clin Oncol* (2014) 32(29):3236–41. doi:10.1200/JCO.2013.53.9452

[6] Brenner DJ. The linear-quadratic model is an appropriate methodology for determining isoeffective doses at large doses per fraction. *Semin Radiat Oncol* (2008) 18(4):234–9. doi:10.1016/j.semradonc.2008.04.004

[7] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science* (2001) 291(5507):1304–51. doi:10.1126/science.1058040

[8] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* (2001) 409(6822):860–921. doi:10.1038/35057062

[9] Tucker SL, Turesson I, Thames HD. Evidence for individual differences in the radiosensitivity of human skin. *Eur J Cancer* (1992) 28A(11):1783–91. doi:10.1016/0959-8049(92)90004-L

[10] Garber K. Oncologists await historic first: a pan-tumor predictive marker, for immunotherapy. *Nat Biotechnol* (2017) 35(4):297–8. doi:10.1038/nbt0417-297a

[11] Mamounas EP, Tang G, Fisher B, Paik S, Shak S, Costantino JP, et al. Association between the 21-gene recurrence score assay and risk of locoregional recurrence in node-negative, estrogen receptor-positive breast cancer: results from NSABP B-14 and NSABP B-20. *J Clin Oncol* (2010) 28(10):1677–83. doi:10.1200/JCO.2009.23.7610

[12] Nie F, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint l2,1-norms minimization. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*. (Vol. 2), Vancouver, BC: Curran Associates Inc (2010). p. 1813–21. 2997098.

[13] Lee S, Kerns S, Ostrer H, Rosenstein B, Deasy JO, Oh JH. Machine learning on a genome-wide association study to predict late genitourinary toxicity following prostate radiotherapy. *Int J Radiat Oncol Biol Phys* (2018) 101(1):128–35. doi:10.1016/j.ijrobp.2018.01.054

Author Profile

Eash Sharma, Senior Undergraduate, Biochemical Engineering and Biotechnology, IIT Delhi

Ashwin Garg, Senior Undergraduate, Biochemical Engineering and Biotechnology, IIT Delhi

Licensed Under Creative Commons Attribution CC BY