Machine Learning and its Role in Medicine

Ben Fickes

Abstract

Machine learning has been used as a powerful tool to diagnose and treat a wide swathe of medical issues. Intense research and problem solving allows for the design and creation of algorithms specifically tailored for individual maladies, all for the end goal of improving patient outcomes. This paper seeks to illuminate and analyze some common threads between the design of these algorithms, as well as ways in which they differ. Following an overview of machine learning (how it works, and how it is specifically applied to the medical field), this paper discusses, across different sources, how machine learning is utilized to excel in ways that traditional human doctors cannot on their own. Then, it compares the accuracy of various algorithms, as well as other means by which algorithms can be judged as successful or unsuccessful. This cumulates with a comparison between what various algorithms prioritize. Accuracy is utilized as a metric for how adept a machine learning algorithm is at predicting the correct result, but other factors, such as malignant tumors detected, in the case of cancer, bear importance as well. There is a tradeoff between more aggressively targeting malignant tumors (bad outcomes) and accuracy, and this tradeoff requires moral consideration, and a personal weighing of tradeoffs. This paper suggests the employment of varied algorithms to provide multiple perspectives, as several doctors would. Through this discussion, it is hoped that the reader will derive a greater understanding of how machine learning algorithms work, how they provide medical breakthroughs, and what aspects of their design are moral considerations.

Key Terms

Machine learning, medicine, cancer, diagnosis

Machine Learning Overview

In recent times, machine learning has become a popular and frequently discussed topic in both computer science and numerous other disciplines. Its applications are still being discovered, yet it has already been used with great effectiveness. With its ability to detect and act upon trends in data that humans may not be able to discover, it has already been utilized to investigate diagnoses and various treatments for medical conditions, from cancer to PTSD. This paper aims to discuss a sampling of the present uses of machine learning in the medical field, analyze their common strengths, and consider ways in which certain techniques differ from each other and consider and prioritize different indicators of success.

First, it is beneficial to discuss an overview of machine learning. Machine learning is a process by which a computer (a “machine,” if you will) analyzes a data set given to it. It analyzes this information, and, based upon it, “learns” how to classify further data. If the machine learning algorithm is supervised, each training data point will be labelled, and the algorithm will categorize further data into these labels. If the algorithm is unsupervised (not given any labels for these data points), it will group results based on whatever trends in the data it can observe (Bansal, Gupta). The computer is often able to discover patterns or classification criteria not obvious to humans, and thus can be used to both automate and improve procedures that normally require humans to analyze complex data and make judgments (Castrounis). This whole process, however, hinges upon the quality of the data “fed” to the machine; if these data points are incorrectly classified, or do not accurately represent the data the algorithm will classify, the computer will learn incorrectly. Since the computer cannot use the same data it was given as a measure of correctness, it is common to design machine learning algorithms to take in one data set and analyze another. Some variants involve methods such as separating the data into subsets and using each subset independently as a learning set and the others to measure correctness (Kotsiantis, 5). Whatever the exact case, machine learning algorithms require a sufficient amount of data divided into training and test sets in order to be effective and be measured.

There are many different practical applications for machine learning, but this paper will focus upon medical uses. These uses, however, are manifold and varied, and it is difficult to compare and contrast many of them in terms of numerical accuracy or effectiveness. Though most of the examples discussed in this paper revolve around cancer research, discussion will primarily analyze trends in how algorithms are created and measured, rather than providing statistics on how effective various algorithms are. (This information can be gathered from Figure 1, produced later in this paper). Before discussing the creation of these algorithms, however, it is beneficial to consider why they are effective.

Use of Machine Learning in Medicine

In terms of accuracy, different machine learning presents breakthroughs in diagnosing and treating medical conditions in several regards. An enormous number of factors, ranging from genetic to demographic, indicate the presence or risk of cancer (Cruz and Wishart). Early detection of cancer is critical to increasing survivability; doing so can increase survivability rates from 15% to up to 90% (“Why is early diagnosis important?”). Thus, analyzing information and making accurate diagnoses as early and effectively as possible saves patient lives (and, inversely, failing to do so costs them). While a human doctor might be able to reason about some of these, it would be overwhelming, or even impossible, for a person to be able to comprehend and contemplate the full scope of factors. Given a large enough data set to learn from, machine learning, therefore, can discover trends that may go unseen by a human analyst and classify tumors as cancerous or benign. Machine learning isn’t just used for visual analysis; among other things, it has been used to analyze diagnose lung afflictions based on sound (Rongling et. al) and detect epileptic seizures based on electrical impulses in the brain (Shoeb and Guttag). The utility of machine learning, however, extends beyond its ability to diagnose beyond the ability of humans. For example, in treatment of mental health issues (Barish et al.) or in the case of the aforementioned seizures, not only can a machine detect the onset of behavior (or other indicators of an issue), but it can also be constantly present and vigilant in a way that people cannot be.

The majority of breakthroughs in machine learning as applied to medical diagnosis and treatment occurs when a better algorithm is developed. But, what constitutes an algorithm that is “better?” One consideration is the amount of data necessary to make the predictions. In general, machine learning algorithms perform better with a greater sized data set, and so algorithms that can perform well on lesser amounts of data are valuable. (To see the role data size has, consider the different partition sizes and corresponding accuracy rates reported by Chen et al. A greater percent of the data set used as training, with all other variables even, results in a higher accuracy.) However, since the end goal of machine learning in medicine is to perform some sort of detection or prediction, in order to improve and save lives, accuracy is more commonly prized. To ascertain the accuracy of a particular algorithm, after being trained on a data set whose nature is known, it is given another known data set and tasked with analyzing it. Whatever percent of the data is correctly labelled is the algorithm's accuracy. In many different studies and proposals for new algorithms, accuracy is reported as the "result" of the algorithm. For example, Ando et al. reported a 93% accuracy rate when using genes identified by a fuzzy neural network to predict the prognosis of lymphoma. Chen et al. analyze various accuracy rates of an algorithm using different numbers of features (i.e. possible factors to consider), declaring a detection rate of greater than 99% for their five-feature subset. Highlighting the importance attributed to accuracy, one study presents a table of accuracies for various machine learning algorithms in detecting cancer, reproduced in Figure 1 (Cho and Won, Table I).

Table Ⅰ. Relevant works on cancer classification

Authors	Dataset	Method		Accuracy [%]
		Feature	Classifier	
Furey *et al.*	Leukemia	Signal to noise ratio	SVM	94.1
	Colon			90.3
Li *et al.* 2000	Leukemia	Model selection with Akaike information criterion and Bayesian information criterion with logistic regression		94.1
Li *et al.* 2001	Lymphoma	Genetic Algorithm	KNN	84.6~
	Colon			94.1~
Ben-Dor *et al.*	Leukemia	All genes, TNoM score	Nearest neighbor	91.6
	Colon			80.6
	Leukemia		SVM with quadratic kernel	94.4
	Colon			74.2
	Leukemia		AdaBoost	95.8
	Colon			72.6
Dudoit *et al.*	Leukemia	The ratio of between-groups to within-groups sum of squares	Nearest neighbor	95.0~
	Lymphoma			95.0~
	Leukemia		Diagonal linear discriminant analysis	95.0~
	Lymphoma			95.0~
	Leukemia		BoostCART	95.0~
	Lymphoma			90.0~
Nguyen *et al.*	Leukemia	Principal component analysis	Logistic discriminant	94.2
	Lymphoma			98.1
	Colon			87.1
	Leukemia		Quadratic discriminant analysis	95.4
	Lymphoma			97.6
	Colon			87.1
	Leukemia	Partial least square	Logistic discriminant	95.9
	Lymphoma			96.9
	Colon			93.5
	Leukemia		Quadratic discriminant analysis	96.4
	Lymphoma			97.4
	Colon			91.9

Figure 1. *Taken from Cho and Won*

Effectiveness Beyond Accuracy

Despite this importance, accuracy does not wholly encompass the effectiveness of an algorithm. Divyaansh et al. point out that, in breast cancer, the data set is skewed. In a given data set of representative samples, there will be a much higher percentage that are benign, as opposed to malignant. This imbalance will affect the training and accuracy reading of the machine learning, making it more prone to produce a "benign" output than a "malignant" one.[1]

[1] To illustrate this point, consider if the training sample were entirely composed of benign samples. An algorithm would not be able to understand the differences between benign and malignant samples without an example of what was malignant. And, if the test sample were composed of only benign samples, an algorithm that aggressively diagnosed samples as benign would be considered

Divyaansh et al. adjust for this change by weighing malignant tumors more highly in importance in their sample training set. The end goal of this is to produce an algorithm that, given a malignant tumor, will correctly identify it. In these researchers' eyes, an algorithm is valuable if it can successfully diagnose the presence of malignant tumors.

This, however, leads to an important difference in what output is prioritized in an algorithm. In reporting accuracy as a primary desired metric, designers of algorithms assert that accuracy of diagnosis is most important.[2] But, in increasing the weight of malignant tumors, Divyaansh et al. are directly contending that correctly identifying malignant tumors as such outweighs correctly identifying benign tumors as such (or that identifying benign tumors as malignant is less problematic than identifying malignant tumors as benign). In a similar vein, Shoeb and Guttag report the results of their machine learning approach firstly in what percent of test seizures were detected, as opposed to overall accuracy. While it may seem like erring on the side of over-aggression toward detecting malignant tumors (or seizures, or any medical issue) yields the most positive results, this is not necessarily the case. Taken to its extreme, consider an algorithm that simply always diagnoses a sample as a malignant tumor. This would have a success rate of 100% when it comes to detecting tumors, but, as it would lead to many false positives (and, likely, unnecessary procedures, tests, and other trials and expenses), it would not achieve a desired result. A more accurate algorithm that could identify even a portion of the benign cases as such would arguably achieve a greater good.

Making a concrete statement as to how heavily negative outcomes should be weighted is difficult; how aggressively an algorithm marks tumors as malignant may have economic impacts that affect patients differently. Part of any medical treatment involves some measure of cost benefit analysis and cost effectiveness analysis (CDC). Theoretically, this is a decision made by each individual patient. But algorithms would bear the responsibility of making this decision for a great deal of patients. When designing an algorithm, then, one would have to bear in mind aspects unique to each malady: how costly different erroneous classifications are, and the demographics affected by it.

Given enough time for machine learning algorithms to become more prevalent in medicine, one possible solution could be the employment of multiple algorithms for diagnosis. Just as a patient might seek the expertise of multiple doctors, the diagnosis by multiple algorithms would allow for a more complete picture. (Because certain factors that influence each algorithm, such as weight placed upon negative outcomes, are known, which algorithms draw which conclusions would be illustrative in and of itself). For example, if both a "regular" and an "aggressive" algorithm diagnose a tissue sample as benign, the patient knows they most likely do not have cancer. If the two conflict, the patient can employ a cost-benefit analysis to decide for themselves how to proceed.

incredibly accurate. This is mechanically similar to the shortcomings many facial recognition data sets suffering from a lack of diversity cause (Vincent).

[2] This is not to say that the cited works necessarily prioritize accuracy over the detection of malignant or "problem" cases. In particular, Chen et al. do discuss in detail how many malignant and benign tumors they detected and misclassified. In the given data set, they detect all of the malignant tumors present. So, reporting accuracy is not necessarily due to not caring about the detection of malignant tumors, but it does stand out as what the authors present foremost to the reader.

Thus, how significantly bad outcomes (if they are the minority of the sample) should be considered will affect how machine learning algorithms behave, but the judgment call is ultimately a moral, not technological, one, and not easily decided.

Conclusion

Machine learning has been used to great effectiveness to diagnose, treat, and predict various maladies. Many researchers have undertaken the task of constantly refining and crafting new algorithms to produce a greater level of specificity and detail in analyzing data, created an improved ability to diagnose and treat medical conditions. However, some aspects of designing machine learning algorithms require moral judgment calls about how to weight various data points and outcomes, for the ultimate purpose of providing algorithmic outcomes that accomplish the most good. This means that machine learning in the medical field, beyond just centering around objective measurements like accuracy, demands a much more subjective aspect to its design.

Acknowledgements

Thank you to the class of Northeastern University's ENGW3315 in providing suggestions, guidance, and feedback concerning the writing of this paper. Your input has helped tremendously.

References

Ando, Tatsuya et al. "Fuzzy Neural Network Applied to Gene Expression Profiling for Predicting the Prognosis of Diffuse Large B-cell Lymphoma." N.p, November 2002. Web. 21 November 2019. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1349-7006.2002.tb01225.x

Bansal, Shubham. "Supervised and Unsupervised learning." *GeeksforGeeks*. N.p, n.d. Web. 21 November 2019. https://www.geeksforgeeks.org/supervised-unsupervised-learning/

Barish, Greg et al. "A Mobile App for Patients and Those Who Care About Them: A Case Study for Veterans with PTSD + Anger." *ACM*. PervasiveHealth'19, May 2019. Web. 21 November 2019. https://dl.acm.org/citation.cfm?id=3329248

Castrounis, Alex. "Machine Learning: An In-Depth Guide - Overview, Goals, Learning Types, and Algorithms." *innoarchitech*. InnoArchiTech, 2019. Web. 21 November 2019. https://www.innoarchitech.com/blog/machine-learning-an-in-depth-non-technical-guide

Chen et al. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis." *ScienceDirect*. Elsevier Ltd, 2011. Web. 21 November 2019. https://www.sciencedirect.com/science/article/pii/S0957417411001400

Cho, Sung-Bae and Won, Hong-Hee. "Machine learning in DNA microarray analysis for cancer classification." *ACM*. Australian Computer Society, 2003. Web. 21 November 2019. https://dl.acm.org/citation.cfm?id=820213

Cruz, Joseph A. and Wishart, David S. "Applications of Machine Learning in Cancer Prediction and Prognosis." N.p, 1 January 2006. Web. 21 November 2019. https://journals.sagepub.com/doi/full/10.1177/117693510600200030#_i9

Divyaansh, Devarriya et al. “Unbalanced breast cancer data classification using novel fitness functions in genetic programming.” *ScienceDirect*. Elsevier Ltd, 2019. Web. 21 November 2019. https://www.sciencedirect.com/science/article/pii/S0957417419305767

“Economic Evaluation Overview.” *Center for Disease Control and Prevention*. N.p, 2 January 2019. Web. 21 November 2019. https://www.cdc.gov/policy/polaris/economics/index.html

Gupta, Mohit. “ML | Types of Learning – Part 2.” *GeeksforGeeks*. N.p, n.d. Web. 21 November 2019. https://www.geeksforgeeks.org/ml-types-learning-part-2/

Kotsiantis, S.B. Supervised Machine Learning. “Supervised Machine Learning: A Review of Classification Techniques.” Informatica 31, 2007. Web. 21 November 2019. https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Rongling, Lang et al. “Graph-based semi-supervised one class support vector machine for detecting abnormal lung sounds.” *ScienceDirect*. Elsevier Inc, 2019. Web. 21 November 2019. https://www.sciencedirect.com/science/article/pii/S0096300319304667

Shoeb, Ali and Guttag, John. “Application of Machine Learning To Epileptic Seizure Detection.” ICML, 2010. Web. 21 November 2019. https://icml.cc/Conferences/2010/papers/493.pdf

Vincent, James. “IBM hopes to fight bias in facial recognition with new diverse dataset.” *The Verge*. N.p, 27 June 2018. Web. 21 November 2019. https://www.theverge.com/2018/6/27/17509400/facial-recognition-bias-ibm-data-training

“Why is early diagnosis important?” Cancer Research UK. N.p, 26 June 2018. Web. 21 November 2019. https://www.cancerresearchuk.org/about-cancer/cancer-symptoms/why-is-early-diagnosis-important