Pratima Kshetry, pkshetry@umd.edu

Machine Learning In Health Informatics And Breast Cancer Diagnosis

Abstract:

The rise and the evolution of big data has completely revolutionized the field of medical science. From telemedicine to diagnosing diseases, the availability and the use of technology has democratized the research procedures in the field of medical science. The availability of clinical and medical data has provided research scientists with the capability to predict an outbreak of a disease or an early diagnosis of a disease.

This paper discusses how Machine Learning algorithms can be used to detect and diagnose breast cancer. I will illustrate the same with an example dataset which is publicly available at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 using the Knn (K neaest neighbor) algorithm which is one of the many machine learning algorithms. The data was donated by the research scientists at the University of Wisconsin. As per the information made available by the scientists (who are also the donors of the given dataset), the features in the dataset are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass which describe characteristics of the cell nuclei present in the image [1]. The study is based on a statistical tool package R.

Keywords: Machine Learning, Breast Cancer, Bio-Informatics

1. Introduction

Machine Learning is a discipline of computer science, which deals with the development of algorithms so that the machines can learn the behavior of the data using the algorithm and make future predictions. According to Andrew Ng, a professor at Stanford University and founder of Coursera, Machine Learning is the process of getting computers to act without being explicitly programmed. The primary goal of machine learning is to identify patterns in data, and then to perform useful inference using those patterns that have been learned [9]. Such inference typically takes the form of classification, in which previously-unseen data are determined as belonging to one of a number of classes; or regression, in

which previously-unseen data are used to predict the behavior of one or more random variables [10]. An example of classification within healthcare informatics is the determination of whether a hospital patient is "physiologically stable" or "physiologically deteriorating" based on their vital signs [11] ; an example of regression is the prediction of a patient's respiration rate based on physiological data acquired from sensors [12].

Since the boom of big data, companies are extensively using machine learning algorithms for various purposes such as identify and filter spam emails (done by Google), fraud detection (by IBM), and product recommendation (by Amazon) and so on. Although there are many such novel applications of Machine Learning, this research paper however focusses on the application of Machine Learning in the field of Health Informatics especially in diagnosis of a disease.

In today's world, we have a huge availability of medical data. The collection of biological data has increased at an unprecedented rate due to improvements of existing technologies as well as the introduction of new ones that made possible the conduction of many large scale experiments [2]. One such example of an extensive collection of biological data is the 1000 Genomes Project which was launched in 2008 to create a complete and detailed catalogue of human genetic variations, which in turn can be used for association studies relating genetic variation to disease [3]. Other examples of such collection of biological data are GenBank, the U.S. NIH genetic sequence database (www.ncbi.nlm.nih.gov) [2]. This data if used carefully and cautiously in combination of expert knowledge can be used to discover and make breakthrough in the human civilization. This expert knowledge is provided by Machine Learning.

The knn algorithm which we will be implementing in our dataset to demonstrate the usage of such machine learning algorithms in the health diagnosis process, is a widely algorithm in data classification. The k-nn permits the classification of a new data by calculating its distance from all the other data points [8]. The proper functioning of the algorithm depends

on the choice of the parameter k which represents the number of neighbors chosen to assign a label (in our case M-Malignant, B- Benign) to the new data element and the choice of the distance [8]. In this paper, we first illustrate what is knn algorithm and how it can be implemented to diagnose a case of breast cancer. The experiment is conducted using the breast cancer dataset provided by the University of Wisconsin.

The paper is organized as follows. The section 1 introduces machine learning and its usage in health field. Section 2 discusses the literature review or related works on how machine learning algorithms have been used to diagnose diseases. Section three introduces our main discussion topic for this paper which is how a machine learning algorithm can be used to automate the diagnosing process of breast cancer. Section 4 discusses our analytics results obtained by implementing the knn algorithm to our breast cancer dataset. Section 5 evaluates the accuracy rate of the algorithm we used to automate the diagnosis process. We also interpret the results obtained in section 5. Section 6 makes a brief discussion of our output results. Finally, we conclude our study in section 7.

2. Related Work:

Over the years, Machine Learning has been extensively used in the prediction of the outbreak of diseases. Machine learning is widely used in bio-informatics and particularly in the diagnosis of breast cancer. Over the years, researchers have focused on devising better algorithms to automate the detection of several chronic diseases such as heart disease.

Rowan Chakoumakos in one of his research paper “Predicting Outbreak Severity through Machine Learning on Disease Outbreak Reports” presents a machine learning technique that a system can utilize to predict the epidemic potential of each disease outbreak report based on a combination of textual analysis of the individual ProMed report and geospatial)data [4].

Likewise, Joseph A. Cruz and David S. Wishart in their research paper "Applications of Machine Learning in Cancer Prediction and Prognosis" discuss how the Machine Learning techniques can be used to detect and diagnose cancer. The authors claim that they were able to diagnose cancer with 80 percent accuracy. They also insist that if the quality of experimental studies (conducted in Health studies) continues to improve, it is likely that the use of machine learning algorithms will become much more commonplace in many clinical and hospital settings.

Similarly, G. Parthiban and S.K.Srivatsa in their research paper "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients" successfully demonstrate the usage of Naïve Bayes algorithm (a machine learning algorithm) and Support Vector Machines (another algorithm) to diagnose heart disease for diabetic patient based on diabetes diagnosis attributes that was available in their dataset.

3. Diagnosing Breast Cancer with Machine Learning:

Breast cancer is considered to be one of the most leading cause of death in women. The early diagnosis of breast cancer involves examination of the breast tissue for abnormal lumps or masses. If a lump is found, an ultrasound-guided breast biopsy is performed to locate a lump or abnormality and remove the tissue for examination under a microscope [6]. If we could use machine learning algorithms to automate the detection of such cancerous cells, it would be a significant game changer for the existing health systems.

The automation of such processes not only improves the efficiency of the rate of detection which will ultimately increase the number of breast cancer survivors, it will also allow physicians and health professionals to spend more time in curing the disease than in detecting the disease. Such automation processes will also help physicians decide whether a biopsy is required or not. According to breastcancer.org, in the United States, about 80 percentage of biopsies are performed without any need. Because of its cost and

complications, it is essential to decide whether biopsy is necessary or not [7]. Thus, this is where the role of machine learning algorithms come into play to help make medical diagnosis procedure efficient and faster.

As we mentioned earlier, Machine Learning algorithm can be used to predict and diagnose diseases. In this paper, we will demonstrate how a machine learning algorithm can be used effectively to automate the process of diagnosing breast cancer. The algorithm we have used in this research paper is the kNN (abbreviation for k-Nearest Neighbor) algorithm. kNN algorithm works by stores all available cases(data points) and classifies new data based on a nearest distance measure [5]. For example, in figure 1.0, we see that there are two classes for a dataset "a" and "o". We insert a new data point "c" and try to identify which class the new data point "c" belongs to. This can be done using the kNN algorithm. The kNN algorithm identifies three nearest neighbors of a new data point "c". In this case we see that the three neighbors of "c" is "a" and two points of "o". Thus the class of the new element "c" is "o".

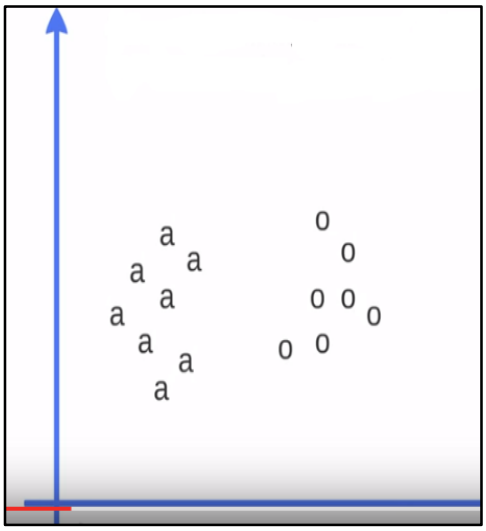


Fig 1.0 Data points of "c"

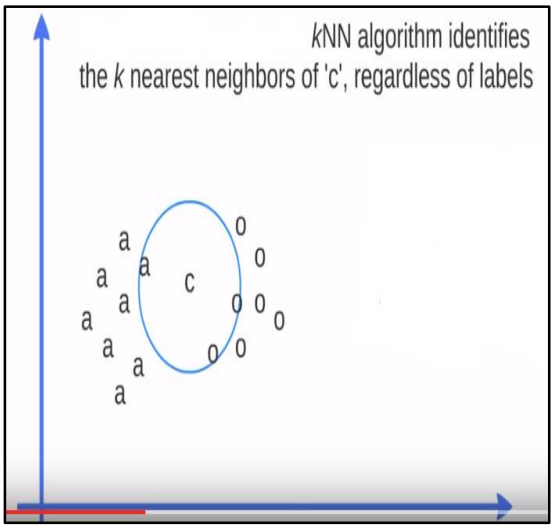


Fig 2.0 Identifying class

The same algorithm can be used to diagnose a breast cancer. The distance between the neighboring points is calculated using the Euclidean distance formula, which can be mathematically as well computationally be computed as:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

4. Data Analysis and Results:

The given dataset consists of 569 observations of cancer biopsies, and 32 variables or features. The first is the identification number, second is the diagnosis (variable) outcome coded as 'M' for malignant and "B" for benign. The other 30 consists of mean, standard error and largest (also the worst case) value for following characteristics of a biopsied cell nuclei:

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1) [1]

Our main objective in this research is to **diagnose** (i.e predict the **Diagnosis** variable) a case as Malignant or Benign based on these features using the machine learning algorithm – Knn.

We have used the statistical package tool R to derive results. As we mentioned earlier, our outcome variable of interest is the diagnosis variable i.e label the case as Malignant (M) or Benign (B). Machine Learning recommends that the any dataset which will be used to make predictions should be split into two datasets – training dataset and the test dataset.

The training dataset is the dataset whose data points are used to generate a learning model or predictive relationship which will be later used to assessed using the data points of the test dataset. In our case, the first 400 observations is used for the training the dataset while the remaining 169 observations is used for the test dataset. Also, an important point to be mentioned here is that while splitting the available dataset into training and test data, we have excluded the "diagnosis" variable because diagnosis is the variable we are trying to predict after the machine learns the pattern of breast cancer diagnosis using the available data.

To implement knn algorithm in our dataset, R provides a package "class" which provides a set of basic R functions for classification. This package provides a simple knn() function call which can be used on the Training dataset and also to build a classifier. A classifier is a mathematical function implemented by the algorithm (in our case – knn) that will map the data observation to a category (in our case Malignant or Benign). The knn() function takes four parameters namely train, test, class and k where:

- Train: is the data frame containing the training dataset
- Test: is the data frame containing the test dataset
- Class (cl) : is a factor vector with the class of each row in training dataset. (in our case Malign or Benign).
- K: is the integer indicating number of nearest neighbors such as k=3 (nearest neighbors).

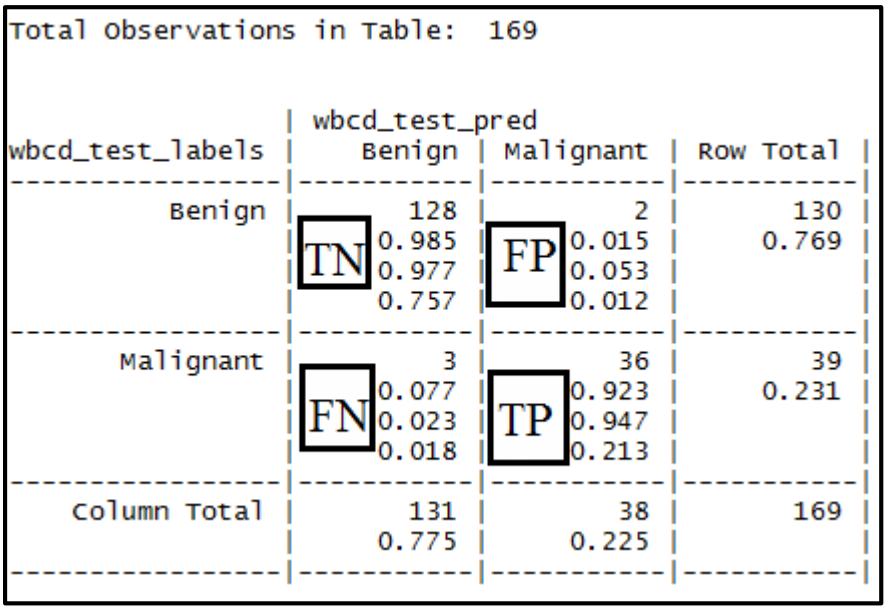
Then, we split the data as training and split. Usually, it is recommended that training data should contain 70-80 percentage of the dataset and remaining as the test data. So in our case, we allot 400 observations in the training data while 169 observations in the test data. Since 400 is the square root of 20, we allocate the value of "k" as 20 i.e the algorithm when tested with the test data will find 20 nearest neighboring observations from the training data to label a case as "Benign" or "Malignant". The following line of code was used to classify the test data:

```
prediction <- knn(train = train_data,
                  test = test_data,
                  cl = train_labels, k=20)
```

The knn() function which is provided by the "class" package in R returns a vector (list) of predicted labels (Malignant or Benign) for each of the data points in the test dataset.

5. Evaluation of the model performance:

This section describes the efficacy of the model we derived in the earlier section. To evaluate the model we have derived using R, I chose to do a cross-tabulation of the output obtained from the prediction results so that we can visually see how well the predicted labels (Malignant or Benign) matched the original labels from the original data. This was achieved using the crosstable() function which is again provided by R. The following screenshot shows the cross-table report of our model:



Total Observations in Table: 169

wbcd_test_labels	wbcd_test_pred: Benign	wbcd_test_pred: Malignant	Row Total
Benign	128	2	130
	TN 0.985	FP 0.015	0.769
	0.977	0.053	
	0.757	0.012	
Malignant	3	36	39
	FN 0.077	TP 0.923	0.231
	0.023	0.947	
	0.018	0.213	
Column Total	131	38	169
	0.775	0.225	

Fig 3. Cross table of prediction model.

The "169" value at the topmost of the table is the total data points we have considered for the test data. The table shows the proportion of values that fall into each of the categories as shown above. In the top left cell, we can see the true negative (TN) results. Of 169 cases,

128 cases were benign and our algorithm correctly labelled them as benign. Also, if we check the bottom right cell (TP or True Positive), 36 malign cases were correctly identified. This means that the algorithmic classifier and the clinically determined labels (Malignant or Benign) both agree that these particular cases were malignant. However, on the top right (cases of False positive (FP)) we see that that 2 benign cases were incorrectly identified as malignant. Such errors are however considered to be least dangerous unlike the cases of False Negative (FN, bottom left). In our case study, 3 of such malignant cases were incorrectly identified as Benign (FN-False Negative) although it was a case of Malignant. Such errors are considered to extremely dangerous as this may lead a doctor to tell his/her patient that she does not have a tumor although in reality, she has a tumor which can even cost her life.

Next, I also calculated the accuracy of this algorithm to correct detect the malign and benign cases. From fig 3.0, we see that total cases = 169, while

Correctly identified case= TN + TP

=128+36 = 164

Thus, the accuracy of the model = (164 / 169) * 100 = **98%** (approx.)

This means that the accuracy of the model derived is 98% (which is a good number) and only 2% of the cases were incorrectly detected by our model.

6. Discussion:

Although our model produced 2% False negative results (which although is not a big number), but such errors can cost someone's life. However, we cannot ignore the 98% accuracy of our model. Thus, perhaps we can try some other more efficient models that will help us reduce the percentage of false negative results to 0 (zero) percentage. The knn algorithm is only one instance of machine learning algorithms. As the field of machine learning algorithm is constantly growing, in future we can expect algorithms that we will certainly yield 0% false negative cases.

Also, the breast cancer is one example of how scientists and medical professionals can automate the process of tumor detection, the same procedure can also be applied to other diseases such as heart disease, brain tumor to name a few.

7. Conclusion:

In this paper, I have highlighted and shown how a machine learning algorithm –Knn can be used to automate the process of breast cancer diagnosis. There are several other machine learning algorithms that can be used for such processes, however I chose to demonstrate the automated diagnosis process using the Knn algorithm.

The Knn algorithm, as we could see appeared to be truly affective producing results with **98%** accuracy. Although the hospitals are not easily convinced to use such automated processes, but with so much of extensive and comprehensive research going on in the field of health informatics, we can expect such automated systems to be a part of the health system in the future.

Such automation of diagnosis process will offer two major benefits. First, the doctor can spend more time in curing the disease than spending a lot of time in diagnosing the disease at the first place. Next, it will help physicians quickly determine whether a further surgery or clinical care is required or not.

References:

1. Available at: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

2. Christos Berberidis "Machine Learning and Data Mining in Bioinformatics". Available at: http://www.academia.edu/1866525/Machine_Learning_and_Data_Mining_in_Bioinformatics

3. Available at: https://en.wikipedia.org/wiki/1000_Genomes_Project#Goals

4. Joseph A. Cruz, David S. Wishart . "Applications of Machine Learning in Cancer Prediction and Prognosis". Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.8502&rep=rep1&type=pdf

5. Available at: http://www.saedsayad.com/k_nearest_neighbors.htm

6. Available at: http://www.radiologyinfo.org/en/info.cfm?pg=breastbius

7. Mahmut Kaya, Oktay Yıldız, Hasan Şakir Bilge ."Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation". Available at: http://www.world-education-center.org/index.php/P-ITCS/article/viewArticle/2642

8. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou. "Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules" . Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.303.6019&rep=rep1&type=pdf

9. David A. Clifton, Jeremy Gibbons, Jim Davies, Lionel Tarassenko. "Machine Learning and Software Engineering in Health Informatics ". Available at: http://www.cs.ox.ac.uk/people/jeremy.gibbons/publications/ml-se-hi.pdf

10. Jeremy Gibbons. Available at: http://www.academia.edu/2959216/Machine_learning_and_software_engineering_in_health_informatics

11. D. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," Journal of Signal Processing Systems, vol. 65, pp. 371–389, 2011.

12. D. Meredith, D. Clifton, P. Charlton, J. Brooks, C. Pugh, and L. Tarassenko, "Photoplethysmographic derivation of respiratory rate: A review of relevant respiratory and circulatory physiology," Journal of Medical Engineering and Technology, vol. 36, no. 1, pp. 60–66, 2012.