Prediction of the risk of cervical cancer based on machine learning techniques

Claudia Cook
BENG 420
Geroge Mason University
Fairfax Virginia
ccook29@masonlive.gmu.edu

***Abstract* — Cancer is group of diseases that involve uncontrolled and abnormal cell growth that can affect different parts of the body. It can have a slow progression but sometimes it can be very aggressive, potentially invading other parts of the body from where it was originally located. It is essential to identify this disease early and give the necessary treatment, but a more effective solution would be preventing it from happening. For this we should investigate the risk factors related to the occurrence of this disease. This project will be more specifically centered on the risks of cervical cancer.**

Keywords — Data mining, machine learning, classification, clustering, artificial neural networks (ANN), K-nearest neighbor (KNN)

I. INTRODUCTION

Cancer is a leading cause of death worldwide, accounting for 8.8 million deaths in 2015. There were about 60.000 detected cases and 30.000 deaths annually up to 2010, with highest incidence in Eastern European countries.

Cervical cancer is a type of cancer that occurs in the cells of the cervix, the lower part of the uterus.
Cervical cancer is caused, 90% of the time, by human papillomavirus (HPV) infection. Even thought the vaccine for this condition was approved in December 2014, cervical cancer remains a significant cause of mortality all around the world. This immunization program is only available in countries with well-organized prevention programs, being many the countries that lack this organization and cant afford it. [1]

Research has shown that there are risk factors involved, such as tobacco use, alcohol use, unhealthy diet, physical inactivity… [2]

When identified early, cancer is more likely to respond more effectively to treatment and can result in a greater probability of surviving. For this reason, the main idea of this project will be the prediction of the risk of cervical cancer based on machine learning techniques.

Is it possible to determine the vulnerability of an individual to cervical cancer based on lifestyle options?

For this purpose, two different data mining techniques will be used.

Data mining can be defined as knowledge discovery; a process of extracting new, useful, valid and potentially understandable patterns from enormous amounts of data. This technique has a great number of applications. Many of them are centered on business and marketing matters, such as sales forecasting, customer loyalty, merchandise planning… [3]. However, its application in healthcare is steadily growing and is becoming essential. It can help diagnose a disease, choose a treatment plan… making the healthcare system cheaper and more effective. There are huge amounts of data generated by healthcare transactions or research studies are too complex and voluminous to be analyzed any other way. [4] This is why we need data mining techniques.

There are 2 papers that analyze the results of the application of different machine learning methods on the dataset that will be used in this project. In the paper *Determining cervical cancer possibility using machine learning methods*; K-nearest neighbor (KNN), multilayer perceptron and Bayes net were used, concluding that KNN gives the best accuracy with 86 nearest neighbors, with a testing [5]. Other papers perform this study using artificial networks, supply vector machine (SVM) and decision trees.

In *Supervised deep learning embedding's for the prediction of cervical cancer diagnosis,* they tackle high-dimensional classification problems by reducing these dimensions, they also use artificial neural network methods. They validate the performance using models such as support vector machines (SVM), KNN and decision trees; concluding that the SVM yields the best results. [6]

II. Materials and methods

In this study, a dataset containing information about cervical cancer patients will be used. The dataset was retrieved from UCI machine learning repository, and was collected at the Hospital Universitario de Caracas, Venezuela. It consists of 36 attributes from 858 patients. The attributes consist on different risks (STDs, contraceptives, different habits…) that in some way affect the probability of this cancer to occur in women. [7]. Some of them are Boolean type, others are integers. There are some missing values from people that didn't want to share determined information, which is represented by question marks.

In this project not all attributes are used. Since this study focuses on the vulnerability of individuals to have cervical cancer based on lifestyle options, the attributes chosen will be the ones that are based on personal choices (contraceptives, smoking…). The other ones, such as age, certain diseases, will be discarded. This leaves us with a total of 10 attributes.

However, the problem of missing values remains. These values have to be replaced. First the amount of missing values for each attribute is counted. It was considered that an attribute is not reliable if it contains more than 20% of missing values; it is non-conclusive. If the percentage is higher than 20%, we will eliminate this attribute altogether. If not, the mean of each attribute will be calculated, the missing value will be replaced with it's corresponding average. This way, the data is altered in the smallest way possible while not loosing information.

To perform this research, I would like to compare two data mining techniques including k-nearest neighbor (KNN) and artificial neural networks (ANN). My hypothesis, based on previous work done on this topic, mainly presupposes that k-nearest neighbor is the best approach for the prediction of this condition based on existent risks, therefore yields smallest error percentage.

A. KNN

This is one of the most straightforward machine learning technique. It is a type of supervised classification learning technique, it needs some sort of training data set to learn from, with attributes and their respective class; thanks to this it will be able to predict the class for different data sets. The idea behind it is straightforward; the program will assign a class to a data point based on the nearest neighbors. It will run through a determined number of nearest neighbors (In this project from 1 to 90), and choose the optimal number of neighbors (optimal k value). It is considered an instance-based technique because it is based directly on training examples, and no previous equations or assumptions.

To calculate the proximity of the training data to the test data we use Euclidean distances as shown below. *d* is the Euclidean distance between the training sample *x* and the testing sample; *i* corresponds to the dimension of the problem or attributes, 10 in total for out dataset.

$$d^j = \sqrt{\sum_{i=1}^{n} (x_i^j - z_i)^2}$$

For validation we calculate the percentage error of each selection of k, this way we will chose the optimal k, which will be the one with the best accuracy.

In this project we will use a training group of 80%, which corresponds to 684. The rest of the dataset (174) will be assigned to the test dataset. We will compute the distances between the different attributes of the test dataset to the training group, predicting using the 90 nearest neighbors. We will assess the accuracy for each prediction based on the nearest neighbor, and choose the one that yields the lowest error.

Some advantages of this method and a reason for choosing it is that it is generally accurate, insensitive to outliers and very flexible to data. Some inconvenient we could encounter is that it is computationally expensive, especially if we have a high number of attributes. It requires time and memory.

B. ANN

Artificial neural networks are modeled in and attempt to imitate the human brain. It is a system composed by many

processing elements opearting in parallel. It is capable of machine learning but also of clustering, pattern recognition. ANN includes a great number of processing units that work together to process information, and like all data ming techniques, extracts useful or at least interpretable information.

The working method is the following: there are initial values for the weights that are learned from training data, and a threshold values that also comes from the training data. The purpose is to find the equation of a linear hyperplane, which is the decision boundary

$$\theta^t x + \theta_o = 0$$

θ^t = weights

θ_o = initial values

The ideal weights will be calculated, uptated sequentially using the following perceptron algorithm.

All weights will be randomly inititalized, then the neuron potential will be computed the following way:

$$a = \sum_{i=0}^{n} \theta_i \, x_i$$

θ_i = weight

x_i = set of features from the training data

Now we predict the class label *(t)* according to the value of a. Lets say for the sake of simplicity, that we assign *1* if a is greater than 0, and *0* if it is smaller.

Next we will compare the values predicted with the actual values. If the prediction corresponds to the true class label, the weight will remain the same. If not, the weight will be uptated every time this occurs, adding or substracting to its anterior value sequentially the following wayd^j

$$\Delta\theta = \propto (y^j - t^j)x_i$$

$$\theta_i = \theta_i + \Delta\theta$$

y^j = the true label

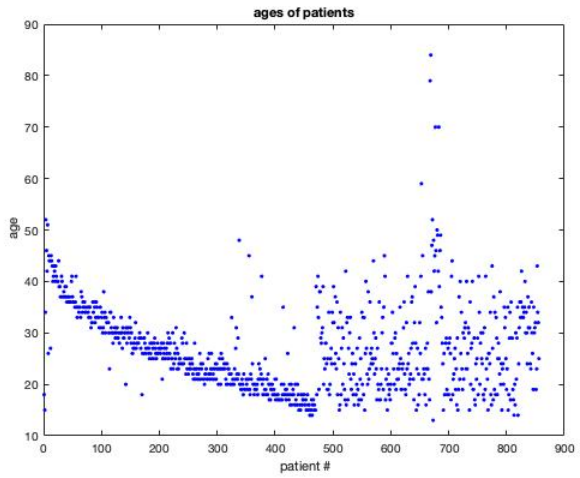
t^j = the predicted label

Now we have the new acorrected value for the weight.

The final corrected weights will be used to make the predcition for the validation data set. The same way as for KNN, we will compare the accuracy of the newly calculated weights by performing the classification again. Then we will compare to the real values and calculate the error. We can compare the accuracy of this method to the accuracy on the KNN with optimal k, an finally choose the best method for the prediction, the one with the highest accuracy.

Some of the advantages of this method is that it is also robust to outliers and corruption to one or more cells doesn't affect too much. They are also able to work with incomplete datasets.

Below is a plot that represents all 856 patients (x axis), and just one attribute (their respective ages).



III. RESULTS

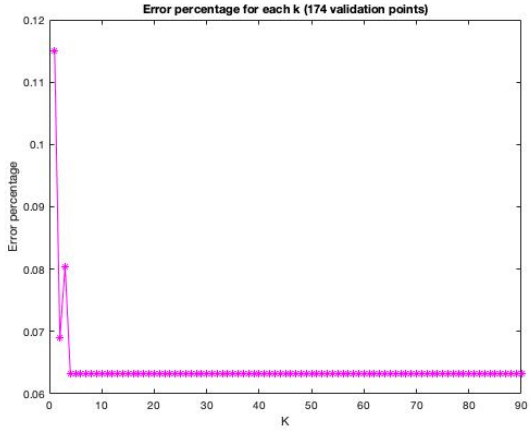
KNN

As mentioned before, we use 684 data points to train the classiffier and 174 for prediction.

As explained in part II, we calculate the percentage of missing value for each attribute. Most of them remained below 3%. The ones with higher percentage of missing values (around 10-12%) correspond to the election of different contraceptive methods (hormonal, IUD). Therefore, all attributes are considered valid for the training of the classifier. The missing values are replaced as explained before.

KNN algorithm is then performed. The predicted classes are compared to the real class labels. Our results indicate that the lowest error is yielded by k = 4, and for k > 4 the error remains the same (6.32%); wich means testing accuracy of 93,68%. Therefore, it is a very effective and fast algorythm. It only needs to identify the class of its 4 nearest neighbors; from that it can predict the class of a new data point correctly.

In the following graph we can see the relationship between the error and number of neigbors used for prediction:



ANN

For this algorithm, the missing values are sorted out the same way as for the KNN algorithm, and the 684 data points with their respective attributes are used to train the classifier.

We call the function *perceptron_neuronN* which will result in the learned weights. Once we have theta, we predicted the values for those same 684 data points.
Since the initial theta is randomly assigned, the final values for the weights vary for a small percentage; therefore the testing accuracy of the perceptron neuron varies, but stays around values 92.98% - 93.57%. This indicates that the data is mostly linearly classifiable. Just a line has been able to separate both classes affectively.
However, we performed the experiment with a different transfer function, the logistic neuron. The testing accuracy of this one is around 98.1%. So this one yields the highest accuracy of them all.

IV. Discussion and conclusion

Going back to our initial hypothesis

Is it possible to determine the vulnerability of an individual to cervical cancer based on lifestyle options?

We can conclude *it is possible*. Both ANN and KNN gave reasonable accuracies on their predictions. We possessed information about lifestyle choices of 856 different subjects, and KNN and linear ANN were able to successfully predict the risk of cervical cancer, with an accuracy of around 93%.

In the paper *Determining cervical cancer possibility by using machine learning methods* [5], knn algorithm is implemented. They also analyze the accuracies for the 90 nearest neighbors. However, maximum accuracy in their paper is reached by k = 86, whereas in this paper the maximum accuracy is reached by k = 4. We presuppose this is because of the increased simplicity in our data. We have only considered 10 attributes instead of 36, which simplifies the problem and eliminates many data points, therefore making the optimal k converge to a solution much faster.

Also, the final accuracy is 93%, whereas the accuracy obtained in the paper was slightly higher, 95.89%. This indicates that maybe additional data (attributes) contribute to a more robust classifier.

ANN linear algorithms weren't used in any of the papers referenced. One of them did use multilayer perceptron algorithms. These yielded a accuracy of 96%.

Therefore, we were surprised to learn that simpler algorithms yielded high accuracy. Even though multilayer perceptron algorithm does yield higher accuracy, linear perceptron follows close behind. However, we are aware that our data attributes were simplified and maybe that is the reason for our high accuracy.

These deep learning methods predicted the diagnosis of different patients with high accuracy, and the next thing that could be done is apply it to other data sets to analyze the robustness and effectiveness of our methods.

Thanks to machine learning methods, we are able to say that these lifestyle choices are significant for the development of this cancer.

References

[1] e. Corusić A, "Cervical cancer as a public health issue--what next? - PubMed - NCBI", *Ncbi.nlm.nih.gov*, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/20432764. [Accessed: 05- Mar- 2019].
[2] Who.int. (2019). *Cancer*. [online] Available at: https://www.who.int/en/news-room/fact-sheets/detail/cancer.
[3]"What is Data Mining? Learn about Definition and Purpose – NGDATA", *NGDATA*, 2019. [Online]. Available: https://www.ngdata.com/what-is-data-mining/. [Accessed: 05- Mar- 2019].
[4] H. Chye Koh and g. tan, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management*, vol. 19, no. 2, p. 1, 2019. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf [Accessed 5 March 2019].
[5] Unlersen, M., Sabanci, K. and Özcan, M. (2017). *Determining Cervical Cancer Possibility by Using Machine Learning Methods*. [online] Available at: https://www.researchgate.net/publication/322233711
[6] K. Fernandes, D. Chicco, J. Cardoso and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis", *PeerJ Computer Science*, vol. 4, p. e154, 2018. Available: 10.7717/peerj-cs.154.
[7] Archive.ics.uci.edu. (2019). *UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29