Breast Cancer Prediction using Data Mining Tool

U. Abirami[1], G. Anantha Jothi[2], S. Ezhilin Freeda[3]

[1,2]*Student, Dept. of Computer Science and Engg., Sri Ramakrishna Engineering College, Coimbatore, India*

[3]*Assistant Professor, Dept. of Computer Science and Engg., Sri Ramakrishna Engg. College, Coimbatore, India*

***Abstract*: Breast Cancer is one of the deadliest diseases that causes majority of deaths among women worldwide. The clustering and classification of the type of breast cancer based on Wisconsin datasets provides accurate prediction of the classes of cancer i.e, Benign and Malignant. This system focuses on applying data mining techniques like K-means for clustering and Regression tree for classification. ANN used for predicting the breast cancer. The goal of this system is to provide thee better accuracy and capable of implementing the algorithms.**

***Keywords*: Diseases, Clustering, Classification, Prediction, ANN.**

1. Introduction

Cancer is one of the most common diseases in the world that results in majority of death. Cancer is caused by uncontrolled growth of cells in any of the tissues or parts of the body. Cancer may occur in any part of the body and may spread to several other parts. Only early detection of cancer at the benign stage and prevention from spreading to other parts in malignant stage could save a person’s life. There are several factors that could affect a person’s predisposition for cancer. Education is an important indicator of socioeconomic status through its association with occupation and life-style factors. A number of studies in developed countries have shown that cancer incidence varies between people with different levels of education [1]. There are several data mining functions such as Concept descriptions, Classification, Prediction, Clustering and Sequence discovery to find the useful patterns. Breast cancer is an uncontrolled growth of breast cells. It refers to a malignant tumor that has from cells in the breast. Breast Cancer constitutes a major public health issue globally with over 1 million new cases diagnosed annually, resulting in over 400,000 annual deaths and about 4.4 million women living with the disease [2]. A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous).

2. Literature survey

Meriem Amrane [3], have compared various classifier algorithms on Wisconsin Breast Cancer diagnosis dataset. The Breast Cancer Dataset (BCD) that they used is donated to the University of California, Irvine (UCI). The classification’s aim is to put each observation in a category that it belongs to. The two machine learning classifiers which are used is Naïve Bayesian Classifier and K nearest neighbor. The purpose is to determine whether a patient has a benign or malignant tumor. The purpose of this article is developing effective machine learning approaches for cancer classification using two classifiers in a data set. The performance of each classifier will be evaluated in terms of accuracy, training process and testing process. In cancer classification, KNN can be used to measure the performance of false positive rates. Naïve Bayesian classifiers are generally used to predict biological, chemical and physiological properties. In cancer classification. On the Wisconsin Breast & KNN, since their target and challenge from breast cancer classification is to build classifiers that are precise and reliable. After an accurate comparison between our algorithms, they noticed that KNN achieved a higher efficiency of 97.51%, however, even NB has a good accuracy at 96.19 %, if the dataset is larger, the KNN’s time for running will increase.

3. Background

Different machine learning techniques are used in our cancer classification.

A. Machine learning approaches

Machine learning is branch of artificial intelligence, ML methods can employ statistics, probabilities, absolute conditionality, Boolean logic, and unconventional optimization strategies to classify patterns or to build prediction models [4]. Machine learning can be divided into two categories: supervised learning (classification) and unsupervised learning. Different Machine learning Techniques have been used in this process.

1) Naïve Bayesian classifier (NBC)

A Bayesian method is a basic result in probabilities and statistics, it can be defined as a framework to model decisions. In NBC, variables are conditionally independent; NBC can be used on data that directly influence each other to determine a model. Bayesian classifiers use Bayes theorem, which is:

$$p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)}$$

- P(h) is the priori probability that event h will occur.

P(d) is the prior probability of the training data. The conditional probability of d when p (d | h) is given.

- P(h | d) is the conditional probability of h when given training data. P (h | d) is the probability of generating instance d given class h. In the equation above Bayesian decision theorem is used to determine whether a given xi belongs to Si where Si represents a class.

$$P(x|S_i)P(S_i) > P(x|S_j)P(S_j)$$

- Si and Sj are two different classes and X belongs to Si.

2) K-Nearest Neighbors (KNN)

The KNN algorithm is used to predict the class or property of data. Given N training vector, suppose we have A and Z as training vectors in this bidimensional features space, we want to classify c which is feature vector [5]. Classifying c depends on its k neighbors, and the majority vote, k is a positive integer, k is generally smaller then 5, if k=1 the class of c is the closest element from the two sets to c.

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

4. Datasets

A. Datasets

The Breast Cancer Dataset (BCD) that we used is donated to the University of California, Irvine (UCI). There are 11 attributes and the first one is ID that we will remove. The nine criterions are as discussed earlier in breast cancer classification section, they are meant to determine if a tumor is benign or malign, the last feature contains a binary value (2 for benign tumor and 4 for malign tumor). The set consists of 699 clinical cases. The initial BCD contains missing data for 16 observations, which limited our dataset to 683 samples.

B. Breast cancer classification (BCC)

To make a good prognostic, breast cancer classification needs nine characteristics which are: 1.determine the layered structures (Clump Thickness); 2. Evaluate the sample size and its consistency (Uniformity of Cell Size); 3. Estimate the equality of cell shapes and identifies marginal variances, because cancer cells tend to vary in shape (Uniformity of Cell Shape); 4.Cancer cells spread all over the organ and normal cells are connected to each other (Marginal Adhesion); 5. Measure of the uniformity, enlarged epithelial cells are a sign of malignancy (Single Epithelial Cell Size); 6. In benign tumors nuclei is not surrounded by cytoplasm (Bare Nuclei); 7. Describes the nucleus texture, in benign cells it has a uniform shape.

The chromatin tends to be coarser in tumors (Bland Chromatin); 8. In normal cells, the nucleolus is usually invisible and very small. In cancer cells, there are more than one nucleoli and it becomes much more prominent, (Normal Nucleoli); 9. Estimate of the number of mitosis that has taken place. The larger the value, the greater is the chance of malignancy (Mitoses) [6]. In order to classify BC, pathologists assigned to each of these characteristics a number from 1 to 10. The likelihood of malignancy needs the nine criteria, even if one of them is very large.

Table 1
Training dataset tuple

Attribute name	**Category**	**Range values**
Id No	Id	-
Clump Thickness	Ordinal	1-10
Uniformity of cell size	Ordinal	1-10
Uniformity of cell shape	Ordinal	1-10
Marginal Adhesion	Ordinal	1-10
Epithelial cell size	Ordinal	1-10
Bare Nuclei	Ordinal	1-10
Bland Chromatin	Ordinal	1-10
Normal Nuclei	Ordinal	1-10
Mitosis	Ordinal	1-10
Cancer	Class	0,1

5. Proposed system

In this system, we data mining techniques like K-means for clustering and Regression tree for classification. ANN used for predicting the breast cancer.

A. Clustering

In clustering process, data is partitioned in to sets of clusters or sub-classes. We have used K-means clustering algorithm. Figure 1 shows that they are clustered as Benign and Malignant. The K-means clustering algorithm works by partitioning n observations in to k sub-classes defined by centroids, where k is chosen before the algorithm starts. The algorithms start with initial estimates for the K centroids, which can either be randomly generated or randomly selected from the data set.

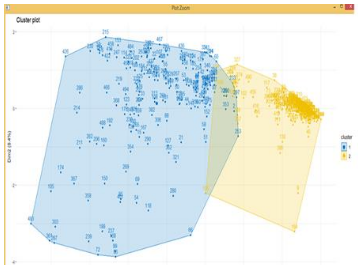


Fig. 1. K-means clustering

6. Classification

Classification is done to partitioning the data into different classes according to some constrains or it classify each item in a dataset into one of predefined set of classes or groups [6].

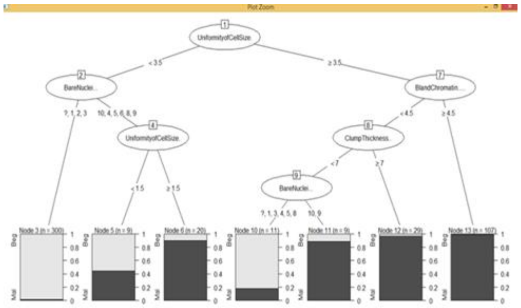
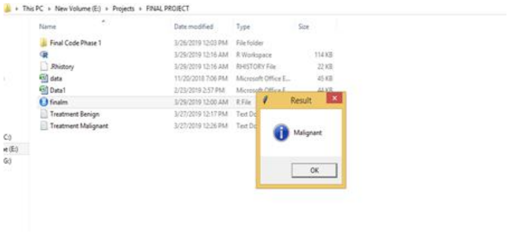


Fig. 2. Regression tree classification

A. ANN

Artificial neural networks are used for prediction. The inputs are compared with the preprocessed data. The resulting predictions will be whether the person has benign or malignant tumor.


This PC › New Volume (E:) › Projects › FINAL PROJECT

Name	Date modified	Type	Size
Final Code Phase 1	3/26/2019 12:03 PM	File folder	
	3/29/2019 12:16 AM	R Workspace	114 KB
.Rhistory	3/29/2019 12:16 AM	RHISTORY File	22 KB
data	11/20/2018 7:06 PM	Microsoft Office E...	45 KB
Data1	2/23/2019 2:57 PM	Microsoft Office E...	41 KB
finalm	3/29/2019 12:00 AM	R File	
Treatment Benign	3/27/2019 12:17 PM	Text Do	
Treatment Malignant	3/27/2019 12:26 PM	Text Do	

Fig. 3. Prediction Model

B. Experimental results

The experimental results of this study using two classifiers are discussed using R-Tool. From the table [2], we can notice that the two algorithms are extremely effective in the diagnosis, all of which shows a high level of accuracy despite in the small dataset. ANN classifiers are ranked first in terms accuracy and duration [7].

1) Existing system

Table 2
Comparison between KNN and NB

Method	Accuracy	Training Process	Testing Process	Total process
KNN	97.51	0.00735	0.001744	0.002479
NB	96.19	0.00759	0.000759	0.001182

2) Proposed system

Table 2
Comparison between Regression tree and ANN

Method	Accuracy	Training process	Test Process	Total Process
Regression Tree	97.79	0.00293	0.00245	0.0053
ANN	99.24	0.00674	0.0033	0.0101

7. Performance evaluation

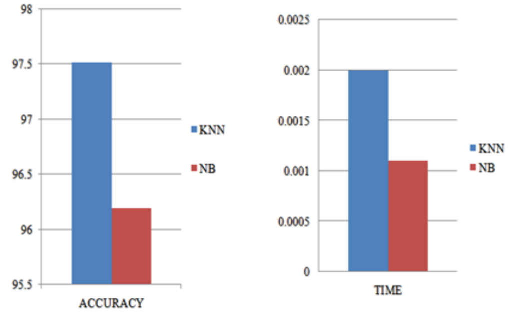


Fig. 4. Accuracy and time comparison for KNN and NB

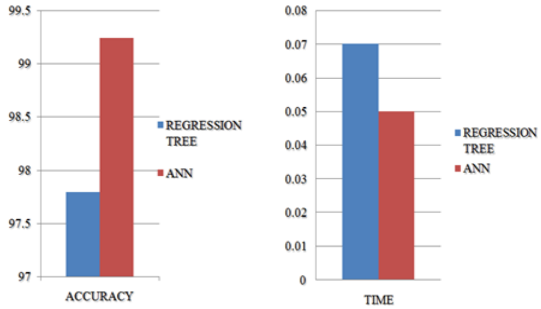


Fig. 5. Accuracy and time comparison for regression tree and ANN

8. Conclusion

In this study, two main algorithms which are Regression Tree and ANN, since our target and challenge is to build classifiers that are precise and reliable. After an accurate comparison between our algorithms, we noticed that ANN achieved a higher efficiency of 99.24%, however even Regression tree has a good accuracy at 97.79%.

9. Future enhancement

This system may play vital role in earlier diagnosis process for different types of cancer and provide effective preventive strategy. In future, the number of attribute covered by the classifier can be increased by increasing the sample size of the training set and hence the development model will be more accurate.

References

[1] P. Ramachandran, N. Girija, T. Bhuvaneswari, “Early Detection and Prevention of Cancer using Data Mining Techniques,” International Journal of Computer Applications, Volume 97, No.13, July 2014.

[2] Akinsola Adeniyi F, Sokunbi M. A, Okikiola F. M, Onadokun I. O, “Data Mining for Breast Cancer Classification,” International Journal of Engineering and Computer Science, Volume 6, Issue 8, pp. 22250-22258, August 2017.

[3] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, 2018, pp. 1-4.

[4] C. Nalini, T. Poovozhi, “Data Mining Classification Technique Applied for Breast Cancer,” International Journal of Pure and Applied Mathematics, Volume 119, No. 12, pp. 10935-10945, 2018.

[5] M. Sugiyama, "Introduction to Statistical Machine Learning "1ed, ed. T. Green: Morgan Kaufmann, 2006.

[6] L. Adi Tarca, V.J.C., X. Chen, R. Romero, S. Ghici, "Machine Learning and its Applications to Biology", PLoS Comput Biol., Vol. 3, pp. 116-122, 2007.

[7] U. Ojha and S. Goel, "A study on prediction of breast cancer recurrence using data mining techniques," *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Noida, 2017, pp. 527-530.