Decision Tree Model in the Diagnosis of Breast Cancer

Liu Yi
Department of Data Science and Technology
Heilongjiang University, Harbin,
Xuefu Road 74#, China
HDliuyi@Outlook.com

Wu Yi*
Department of Data Science and Technology
Heilongjiang University, Harbin,
Xuefu Road 74#, China
1352900831@qq.com
*Correspondent Author

***Abstract*—Breast cancer is the second leading cause of cancer death in women. At the same time, it is one of the most curable cancer if it could be diagnosed early. More and more researchers have confirmed that the decision tree model has a good ability to accurately diagnose. This paper presents a diagnostic method for breast cancer based on the decision tree model combined with feature selection. Experiments were conducted on different training test divisions of the Wisconsin Breast Cancer Data Set (WBCD), a common method used by researchers to diagnose breast cancer based on machine learning methods. In this paper, in order to reduce the complexity of the decision tree model, this paper proposed to delete some highly relevant features of ... After data correlation and independence tests, it finally chosed the tumor thickness, cell shape consistency, single epithelial cell size and mitosis as a subset of the decision tree model. Experimental results show that the classification accuracy (94.3%) significantly outperforms the state-of-the-art method with respect to a variety of metrics.**

Keywords-Breast cancer diagnosis; Decision-tree; Feature selection

I. INTRODUCTION

Cancer is a disease in which cells grow, divide and reproduce uncontrollably. In general, cancer is named after the body part of its lesion; therefore, breast cancer refers to the presence of cells in breast tissue for unlimited reproduction [1]. A group of rapidly dividing cells may form lumps or lots of noxious tissue, and these lumps are called tumors. Tumors can be cancerous (malignant) or non-cancerous (benign). Malignant tumors penetrate and destroy healthy body tissues.

Breast cancer is a malignant tumor that occurs in the epithelial tissue of the breast, and 99% of them occur in women[7].It is reported that breast cancer, after lung cancer, gastric cancer, liver cancer, esophageal cancer and colorectal cancer, has become the sixth leading cause of cancer death in Chinese women[10].

The application of feature classification in medical diagnosis is gradually maturing. There is no doubt that the collection of patient data and the assessment of experts are the most important factors in diagnosis [2]. However, feature classification systems and different AI techniques can also be of great help to experts.

The decision tree model is composed of decision points, strategy points (event points) and the results of the tree structure, generally used in the sequence decision. Usually the maximum return expectations or the minimum expected cost as the decision criteria, through graphical solutions to the effectiveness of various programs under different conditions, and then make the best decisions by comparison. Among them, the advantages of the decision tree model are as follows: First, the Shallow decision tree is visually intuitive and easy to explain. Second, the data structure and distribution need not make any assumptions. Third, you can capture the interaction between variables (Interaction).

When using the decision tree model for prediction, we should remember that the best input subset of features plays a crucial role in building predictive models with high prediction accuracy and high stability [3].Because the choice of feature subset affects the complexity and accuracy of the decision tree model. Feature selection is an important issue in establishing a feature classification system. Its purpose is to identify important features and eliminate irrelevant features so as to establish a good learning model. Therefore, without using the existing dimensionality reduction techniques, we strongly hope that the information contained in the dataset can be used to find the optimal subset of features. We improve the predictive accuracy of the model by analyzing data correlation and using Pearson's chi-square test for independence between features and labels. From a medical point of view, this is to determine the most important factor that affects the patient's diagnosis (benign or malignant) [4].

II. METHODOLOGY AND EXPERIMENTS

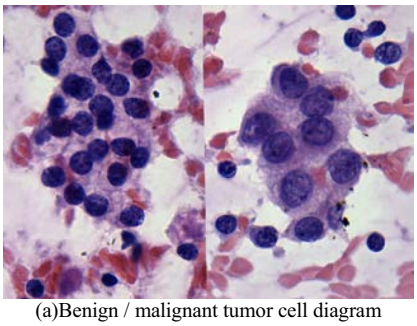
A. Breast cancer dataset

We used the breast cancer data set from the UCI machine learning library in our experiments. This dataset is generally suitable for researchers to use machine learning methods for breast cancer prediction diagnosis, so it is suitable for the performance testing of our models and other machine learning methods [5]. This dataset contains a total of 699 samples, removing a total of 683 samples of unqualified data containing "?"

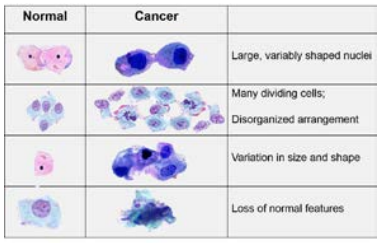
978-1-5386-5784-3/17/31.00 ©2017 IEEE
DOI 10.1109/ICCTEC.2017.00046

B. Feature Selection

According to the characteristics of the data set, we classify them into the following nine categories: Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses.



(a)Benign / malignant tumor cell diagram



(b)Differentiation of mitosis between normal and cancer cells

Fig.1 Data set characteristic diagram

C. Data preprocessing

The data in the dataset has more dimensions, which is not good for us to analyze the data. We take the IsoMap isometric mapping algorithm for dimensionality reduction. It can be seen from the figure that the first few dimensions represent all the data.

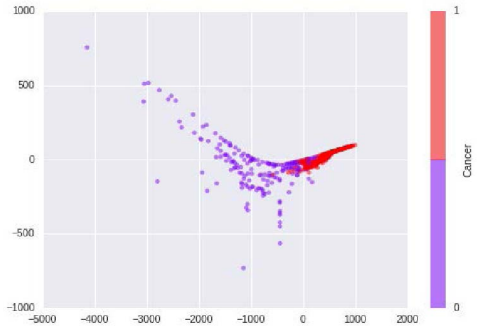


Fig.2 IsoMap Dimension reduction diagram

D. Data correlation analysis

We performed a correlation analysis of 9 features except ID, output the correlation matrix and obtained the correlation coefficient between each feature. As can be seen from the figure, there is a large correlation between features such as "cell size uniformity" and "cell shape uniformity". Their correlation coefficient is greater than 0.9, and they are considered highly relevant.

TABLE 1 DATA CORRELATION DIAGRAM

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bland Chromatin	Normal Nucleoli	Mitoses
Clump Thickness	**1**	**0.90722823**	**0.70637695**	**0.75354402**	**0.69170875**	**0.75555916**	**0.71934604**	**0.4607547**
Uniformity of Cell Size	**0.907228 23**	**1**	**0.68594806**	**0.72246241**	**0.71387755**	**0.7353435**	**0.71796341**	**0.44125758**
Uniformity of Cell Shape	**0.706916 95**	**0.68594806**	**1**	**0.59454777**	**0.67064829**	**0.66856706**	**0.60312106**	**0.4805833**
Marginal Adhesion	**0.753544 02**	**0.72246241**	**0.59454777**	**1**	**0.58571613**	**0.6181279**	**0.6289264**	**0.4805833**
Single Epithelial Cell Size	**0.691708 75**	**0.71387755**	**0.67064829**	**0.58571613**	**1**	**0.68061486**	**0.5842802**	**0.33920144**
Bland Chromatin	**0.755559 16**	**0.735435**	**0.66859706**	**0.6181279**	**0.68061486**	**1**	**0.66560153**	**0.34601089**
Normal Nucleoli	**0.719346 04**	**0.71796341**	**0.60312106**	**0.6289264**	**0.5842802**	**0.66560153**	**1**	**0.43375727**
Mitoses	**0.460754 7**	**0.44125758**	**0.4805833**	**0.4805833**	**0.33921044**	**0.34601089**	**0.43375727**	**1**

In order to reduce the complexity of the decision tree model and improve its accuracy, we decided to remove the highly correlated features from the remaining features to better represent the characteristics of the patient's tumor cells. For features with greater relevance, we will retain the maximum. Finally, we selected four characteristics as a subset of the characteristics of the decision tree model based on the correlation between the data: Clump Thickness, Uniformity of Cell Shape, Single Epithelial Size, Mitoses.

E. Data independence test

Pearson chi-square test was used to verify the independence of features and tags. Test results shown in Figure 3. It can be seen from the figure P value is small, indicating that these features are very relevant label.

Therefore, the characteristics of the choice can be a good judge of malignant cells.

TABLE 2 DATA INDEPENDENCE DIAGRAM

	Clump Thickness	Uniformity of Cell Shape	Single Epitelial Cell Size	Mitoses
Degree of freedom	**9**	**9**	**9**	**8**
Statistic	**378.0616**	**523.0709704**	**447.861175**	**191.9681974**
PValue	**0**	**0**	**0**	**0**

III. DECISION TREE MODEL

We construct the decision tree model from the four selected features and make a 0-1 judgment on each feature so as to achieve a relatively good model with fewer features.

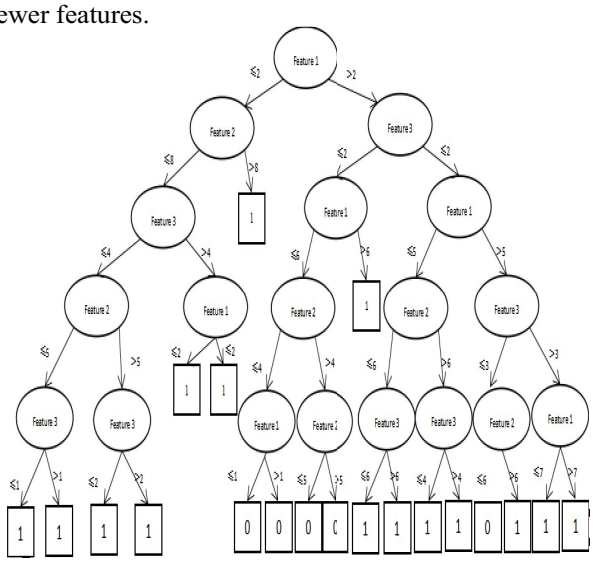


Fig.3 Decision tree model diagram

Feature 1: Uniformity of Cell Shape; Feature 2:Clump Thickness; Feature 3:Single Epithelial Cell Size

IV. MODEL EVALUATION

In order to evaluate the prediction performance of Decision-tree classifier, we define and compute the classification accuracy, sensitivity, specificity and ROC curves, respectively. The formulations are as follows

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} * 100\%$$

$$Sensitivity = \frac{TP}{TP + FN} * 100\%$$

$$Specificity = \frac{TN}{TN + FP} * 100\%$$

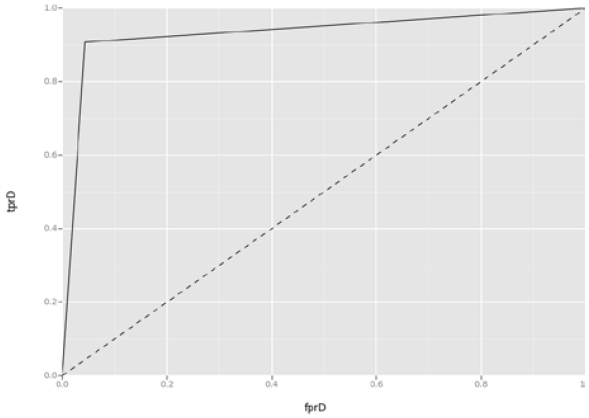


Fig.4 ROC curve [6]

V. CONCLUSION

A. Experimental results

To assess the effectiveness of the proposed method, we performed an experiment in the Wisconsin breast cancer database. Choose a subset of features that are valuable by examining the relevance and independence of the data. Figure 3 shows a decision tree model based on feature selection. Table 1 shows the correlation coefficient results of the test data for eight subsets. Among the eight subsets, the correlation between the subsets is high.

We give an independent test of the feature subset (Table 2) and also give the ROC curve of the decision tree model (Figure 4). Calculating the area under the ROC curve (AUC) can be used to evaluate the classifier performance for different training / testing partitions. The larger area of AUC means that the classifier classification performance is better.

As can be seen from the confusion matrix, false positives and false negatives decrease with increasing training set size. Through the test results, we can see that the accuracy of the decision tree model is 94.3%, which belongs to the model with high accuracy and has certain experimental value. Of the 123 benign tumors, 6 were misdiagnosed and the misdiagnosis rate of malignant tumors was 4.8%.

B. Future work

This work explores a new breast cancer diagnosis model - the decision tree model. Different WBCD experiments show that decision trees perform well in differentiating benign breast from malignant breast tumors [8].

As the results show, the proposed model achieves a high classification accuracy (94.3%) for the selected subset of the four features. At the same time, comparative experiments on the top four related features and all nine features are carried out.

The results show that the four features identified by the decision tree are superior to other feature subsets in terms of highest classification accuracy and average classification accuracy. In addition, combining the features of the breast tumor classification (ie, "tumor thickness", "uniformity of cell shape", "single epithelial cell size" and "mitosis") by the correlation analysis and the independence test is beneficial to the disease the diagnosis. This means that these four characteristics warrant doctors' close attention during their diagnosis.

We believe that this approach (decision tree) has shown promising results in the classification of breast cancer to ensure that physicians make very accurate diagnostic decisions. Future surveys will place great emphasis on evaluating the proposed decision trees in other larger breast cancer datasets [9].

ACKNOWLEDGMENT

This work is supported by the Nature Science Foundation of Heilongjiang Province with the grant number: F2016035. The author would like to thank for their support.

REFERENCES

[1] Jiang G L. Support vector machine Feature selection algorithm based on modified rough set attribute reduction algorithm [J]. Microcomputer Information, 2010, 26(27):192-194.

[2] Akay M F. Support vector machines combined with feature selection for breast cancer diagnosis [J]. Expert Systems with Applications, 2009, 36(2):3240-3247.

[3] Abonyi, J., & Szeifert, F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters, 14*(24), 2195–2207.

[4] Finally. Weight Optimization in Recurrent Neural Networks with Hybrid Metaheuristic Cuckoo Search Techniques for Data Classification [J]. Mathematical Problems in Engineering, 2015,(2015-10-5), 2015, 2015(4):1-12.

[5] Polat K, G 체 ne 힊 S. Breast cancer diagnosis using least square support vector machine [J]. Digital Signal Processing, 2007, 17(4):694-701.

[6] Jaganathan P, Rajkumar N, Nagalakshmi R. A Kernel Based Feature Selection Method Used in the Diagnosis of Wisconsin Breast Cancer Dataset [M]. Advances in Computing and Communications. Springer Berlin Heidelberg, 2011:683-690.

[7] Marcano-Cedeño A, Buendía-Buendía F S, Andina D. Breast Cancer Classification Applying Artificial Metaplasticity[C]// International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer, Berlin, Heidelberg, 2009:48-54.

[8] Ahan S, Polat K, Kodaz H, et al. A new hybrid method based on fuzzy-artificial immune system and k k mathContainer Loading Mathjax -nn algorithm for breast cancer diagnosis[J]. Computers in Biology & Medicine, 2007, 37(3):415-423.

[9] Penareyes C A, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis.[J]. Artificial Intelligence in Medicine, 1999, 17(2):131-155.

[10] Qi, L., & Min, W. (2007). Fuzzy Optimization Control of the Temperature for the Heating Process in Coke Oven Based on Co-evolution. *Control Conference, 2007. CCC* (Vol.27, pp.420-424). IEEE.