

Prediction of Anorexia Risk Based on Decision Tree

Jinyu Xie*

School of Data Science, The Chinese University of Hong Kong (Shenzhen) Shenzhen, China

121040075@link.cuhk.edu.cn

Abstract. Anorexia is a kind of eating disorder caused by the cognitive distortions of patients. Under the condition, patients' physical functions can be impaired by extreme weight control behaviors. The prediction of anorexia can help patients with early detection and prediction of the risk of disease, then decrease the damage to the body and the psychological burden. In recent years, machine learning models achieve promising performances in prediction problems. In this paper, after trying on the Adaboost model, Extratree model, and decision tree model, this research decides to use the decision tree model. It is suitable for Boolean-type data classification to do data analysis and hence achieves outstanding performances. By building a decision tree model and referring to the correlation between different features, this research concludes data training. People who have the problem with liver firm and spleen palpable are more likely to gain anorexia. Fatigue and malaise have little relationship with anorexia.

Keywords: machine learning, anorexia prediction, decision tree.

1. Introduction

Anorexia is one of the eating disorders caused by a distorted perception of weight. People with the condition often take extreme measures to control their weight, such as vomiting after meals, drug abuse, and excessive exercise [1]. Predicting anorexia in advance can prevent patients from the physical impairment and psychological damage which result from taking extreme measures to control their weight [2].

In recent decades, machine learning models achieve promising performances in many research fields [3,4]. It could accurately predict the emergence of a specific disease after being trained with a collected and finely annotated dataset. Compared with training medical experts for diagnosis, machine learning models are much more cost-effective. They could be trained quickly and objectively to classify input samples according to given features. Moreover, it is more suitable to be used to screen the potential patients of anorexia, so that the people with high risk could be noticed in advance and seek doctors for help [5,6].

Recently, many researches focus on the field of prediction of anorexia, since its great importance in keeping healthy. Sayanta, Jandhyala, and Tanmay [7] took research on predicting anorexia through people's social media. Low prealbumin was found as one of the predictors of anorexia by Gaudiani and his partners [8]. RNN-LSTM and SVM classification [9] were once used to predict anorexia through emotional classification.

This research analyses physical fatigue and liver condition by decision tree model to early predict anorexia. The decision tree model belongs to machine learning. Compared with predicting and diagnosing anorexia, the decision tree model can quickly analyze a large amount of data and then come to the conclusion, which for inexperienced medical staff has a good reference value. At the same time, machine learning can combine many factors which may contribute to anorexia together, helping medical workers differentiate between anorexia and general poor appetite or excessive weight loss based on these factors combined with patient characteristics.

2. Method

2.1. Dataset

In this research, the data set is provided by Gail Gong [10]. The data set contains 20 different factors which may lead to anorexia in all. This research chooses 2 features of body condition(“Fatigue”, “Malaise”) and 3 features of liver function(“Liver big”, “Liver firm”, “Spleen palpable”). There are 154 groups of data in whole including 12 groups involving null values.

2.2. Data Pre-processing

Since the proportion of missing values in most groups containing missing values exceeds 50%, this research considers such groups to be of no research value. So it is decided to remove the groups which contain missing values instead of filling them. Also as the distribution of data is not balanced and the whole number of data groups is small, this research decided to oversample the train data by SMOTE. Previous training data has 100 groups, and test data has 43 groups, after oversampling the training data increased to 165 groups, and the number of test data groups remains the same.

2.3. Model Construction

In this research, the Adaboost model and Extratree model are first used in the model trial stage. Adaboost is a model that assigns different weights to different data according to the importance of the data and constantly adjusts the weights of different data based on the previous classification effect. The main idea of the Adaboost model is an iterative algorithm, which adds a new weak classifier in each round. Wrong samples of the previous classifier will be used to train the next classifier and keeps iterating until it reaches a predetermined small enough error rate.

The Extratree model is constructed by many decision trees which use the same training data. Every decision tree applied the same origin training data rather than randomly sampled data. Also, each branch of the decision tree is determined by a random chosen feature, without following any principles.

Decision tree is a model in machine learning, this is also the main model used in this research. The decision tree is a hierarchical tree structure, which is mainly composed of the root node and the child nodes on the branch. Each feature node has a judgment condition, and the result can be divided into two parts. The principle of decision tree is to judge each feature from the root node according to conditional probability, and allocate the results to different child nodes. Information gain and Gini impurity are two methods for selecting the best properties on the node. Information gain represents the difference in entropy before and after splitting a given attribute. The method of computing entropy is shown below:

$$\text{Entropy}(S) = - \sum_{c \in C} p(c) \log_2 P(c) \quad (1)$$

Among the equation, S represents the data sets of information entropy, c represents the class in set, p(c) represents the proportion of data points belonging to class c to the total data points in the data set. The equation of information gain is shown below:

Information Gain(S, a) =

$$\text{Entropy}(S) - \sum_{v \in \text{values}(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (2)$$

Among the equation, a represents a specific attribute or class label, information entropy (S) is the entropy of the data set, Information entropy (Sv) is the result of entropy of the data set, $|S_v|/|S|$ is the proportion of the value of Sv and the value of data set. Gino impurity is the probability of wrong classification when labeling random data point.

The advance of the decision tree model is this model can handle the classification results of irrelevant features. In this research, physical fatigue has no relationship with a liver condition. And since each node in the decision tree has two child nodes, it is suitable for Boolean-type data

classification, and the data type in this research is Boolean. So this research decided to use the decision tree model to analyze the end.

When applying the model, this research first divides the data set into two parts. The first part contains 165 groups of training data with 5 data values which represent 5 features respectively in each group. The second part involves 43 groups of testing data which also have 5 values corresponding to 5 features in every group. The data split results are demonstrated in Table 1. The purpose of classification is to obtain the parameters of the model firstly through the training data, and then the accuracy of the model is tested by testing data, to determine whether the model is suitable or not.

Table 1. The number of training data and testing data

	train	test
Origin	100	43
After	165	43

After the classification of the data set, this research uses Adaboost, model, Extratree model, and decision tree model respectively to train and test data. Finally out put the classification results according to the input data.

2.4. Evaluation Matrix

This research is a classification problem. Classification 0 represents the people who do not have anorexia, classification 1 represents the anorexia patients. There are three model evaluation criteria, prediction, recall, f1-score, and accuracy. Prediction refers to the probability that the sample predicted as positive is actually positive. Recall refers to the probability that the sample is actually positive and also predicted positive. F1-score is the harmonic value of precision and recall. Accuracy refers to the percentage of the total sample that predicts the correct result.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 = \frac{2*Recall*Precision}{Recall+Precision} \quad (5)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

3. Result

3.1. Construct Correlation

In order to see the relationship between different features more directly and effectively, this research adopts the correlation chart in data visualization. Figure 1 below shows the correlation between every two features. The deeper the color is, the lower correlation the two features have. The lighter the color is, the higher correlation the two features have.

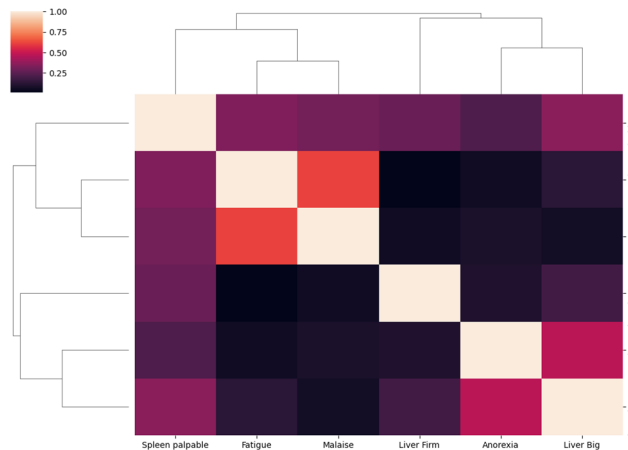


Figure 1 Correlation between different features

From the correlation chart, it can be seen that the highest correlation with anorexia is liver firm and spleen palpable, while physical fatigue and malaise are not highly correlated with anorexia.

3.2. Result of Classification Model

Tabel 2 shows from the Extratree model to the decision tree model, every evaluation standard of every classification and the whole accuracy are increasing. Finally, the Precision, recall, and accuracy of the decision tree are all the highest among the three, the results are more reliable. Among them, the recall of classification 0 and precision of classification 1 reach 1.

Tabel 2 Results of Models

	Precision (0/1)	Recall (0/1)	F1-score (0/1)	Accuracy
Extratree	0.80/0.92	0.95/0.73	0.87/0.81	0.85
Adaboost	0.81/1.00	1.00/0.73	0.89/0.84	0.87
Decision Tree	0.83/1.00	1.00/0.76	0.90/0.86	0.89

Finally, the ROC curve, illustrated in Figure 2, is used to judge whether the model is valuable or not. The ROC curve is an image formed by combining FPR and TPR of the same model in ROC space. The TPR is also the recall value.

$$TPR = \frac{TP}{TP+FN} \quad (7)$$

$$FPR = \frac{FP}{FP+TN} \quad (8)$$

According to the ROC curve, the value of AUC is 0.54 exceeding the median of 0.5 which is better than the guess. So the model is valuable.

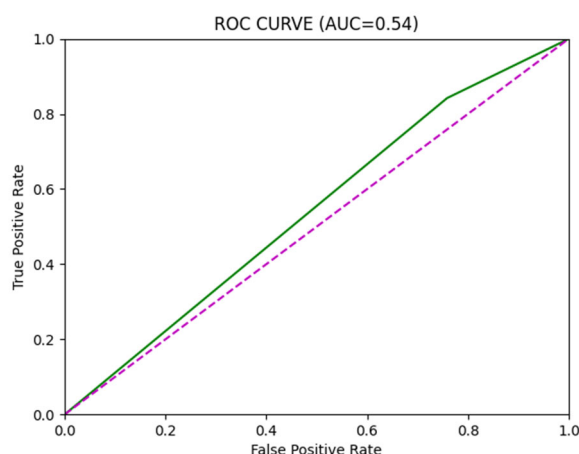


Figure 2. Correlation between different features

4. Conclusion

This research uses the decision tree model to predict anorexia through physical fatigue and liver condition, and the accuracy of the model reached 0.89. The conclusion is liver firm and spleen palpable are more likely to have anorexia while fatigue and malaise are not strongly associated with having anorexia. In the real world, if people have these types of liver problems in physical examinations and at the same time people have dietary abnormalities, then people can predict anorexia rather than suspect some other diseases. Compared with previous research, this research has an increase in recall value, from 0.56 to 0.76. Moreover, the evaluation criteria of the model are more comprehensive, extending from the mean value and standard deviation of previous studies to correlation, accuracy, recall, f1-score, and accuracy. However, the research also has problems such as too little data and not enough training models. Future studies can collect more data from multiple sources and analyze the data obtained from different sources to test the accuracy of the model and determine whether the model is compatible. Other models can be used to improve the accuracy as well.

References

- [1] Mehler, P.S., "Diagnosis and care of patients with anorexia nervosa in primary care settings," *Annals of Internal Medicine*, vol. 134(11), pp.1048-1059, Jun 2001.
- [2] Aguilera J, Farías DI, Ortega-Mendoza RM, Montes-y-Gómez M. "Depression and anorexia detection in social media as a one-class classification problem," *Applied Intelligence*, vol. 51, pp. 6088-103, Aug 2021.
- [3] Mahesh B. "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, pp. 381-386, Jan 2020.
- [4] Bell J. "What is machine learning?," *Machine Learning and the City: Applications in Architecture and Urban Design*, vol 31, pp. 207-216, May 2022.
- [5] Jordan MI, Mitchell TM. "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349(6245), pp. 255-260, Jul 2015.
- [6] Alzubi J, Nayyar A, Kumar A. "Machine learning from theory to algorithms: an overview," *InJournal of physics: conference series*, vol. 1142, pp. 012012, Nov 2018.
- [7] Paul S, Jandhyala SK, Basu T. "Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks," *In CLEF (Working notes)*, vol. 2125, pp. 1-15, Sep 2018.
- [8] Gaudiani JL, Sabel AL, Mehler PS. "Low prealbumin is a significant predictor of medical complications in severe anorexia nervosa," *International Journal of Eating Disorders*, vol. 47(2), pp. 148-156, Mar 2014.
- [9] Ranganathan A, Haritha A, Thenmozhi D, Aravindan C. "Early Detection of Anorexia using RNN-LSTM and SVM Classifiers," *In CLEF (Working Notes)*, vol. 2380, pp. 1-11, Sep 2019.

- [10] Amarnath B, Balamurugan S, Alias A. "Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset," Journal of Engineering Science and Technology, vol. 11(11), pp. 1639-1646, Nov 2016.