A Classification Diagnosis of Cervical Cancer Medical Data Based on Various Artificial Neural Networks

Yong Qi [a], Zhijian Zhao [b], Lizeqing Zhang [c], Haozhe Liu [d] and Kai Lei [e]

Shaanxi University of Science and Technology, Shaanxi, 710021, China.

[a] qiyong@sust.edu.cn
[b] corresponding author: messhiro@gmail.com
[c]18729321203@163.com
[d]liuhaozhest@gmail.com
[e] LeiKai66666@gmail.com

Abstract. This paper mainly proposed to identify and classify the medical data of cervical cancer using neural network models such as SVM, FNN, KNN and so on. The computer recognition algorithm can overcome the deficiency that artificial identification tends to be affected by cognitive ability, subjective experience and fatigue degree. The model trained under various neural networks with the experts' manual marked data will obtain a more precise result in the identification process of cervical cancer medical data.

Key words: health care; cervical cancer; machine learning; deep learning; regression.

INTRODUCTION

Although cervical cancer is the most preventable type of cancer, it kills about 300,000 women worldwide each year. Cervical cancer is diagnosed between 35 and 54 years old and about 20% of women over the age of 65. The median age of women diagnosed with it is 48. About 15% of women suffer from cervical cancer between the ages of 20 and 30. Meanwhile, Cervical cancer is extremely rare among women under the age of 20. Many young women are infected with multiple types of human papillomavirus, which is the leading cause of increased risk of cervical cancer. At present, cervical cancer is diagnosed through cytology, colposcopy, uterine biopsy, cervical conization resection, fluorescein examination, fluorescence microscopy, cervical local rapid diagnosis etc.

Artificial intelligence has been used to support medical treatment for many years. Shi, Xue, et al. proposed three kinds of machine learning method (support vector machine, artificial neural network and deep learning), which performed well in tumor diagnosis and prediction. SVM and ANN have become the indispensable part of the researchers' building model. Cao, Ying, et al established a diagnostic prediction model based on BP neural network, logistic regression and stochastic forest algorithm. They compared the diagnostic value of the three models for prostate cancer, and eventually verified that all three models had higher diagnostic validity.

This paper attempts to classify and diagnose a large number of medical data through SVM, FNN, KNN and other artificial neural network structures. We promote a diagnosis method combining traditional machine learning and deep learning. which can greatly simplify the physician's workload and further improve the accuracy of diagnostic work.

EXPERIMENTS

Our team used a dataset of biopsy cervical cancer risk factor for experiments. The dataset is a CSV format file with 858 records. The attribute information of the dataset contains the data of 35 related test samples as follows: age, the number of sexual partners, having sexual intercourse or not, pregnant or not, smoking or not, smoking years, the amount of cigarettes per year, whether to use hormonal contraceptives, whether to use intrauterine device, years after placing intrauterine device, having sexually transmitted diseases or not , the number of people having sexually transmitted diseases, the presence of condyloma, the presence of cervical condyloma acuminatum, the presence of vaginal condyloma acuminatum, the presence of vulva - perineal acuteness wet wart, the presence of syphilis, the presence of pelvic inflammatory disease, the presence of genital herpes, and the presence of contagious soft wart, the presence of HIV/AIDS, the presence of hepatitis b, the presence of HPV, diagnostic quantity, the time since the first diagnosis , the time since the last diagnosis, Dx: Cancer, DX: CIN, DX: HPV, DX, Hinselmann, Schiller, Citology.

The label indicates whether the test sample has cervical cancer or not.

Our team used machine learning and deep learning algorithms to achieve automatic classification and identification of cervical cancer after network training through the relevant feature attributes.

In the first step, we preprocessed and normalized the sample data, and then transformed the data into a matrix; after that we did data cleaning work, label settling and adjustment on data format.

In the second step, we first trained the preprocessed datasets through machine learning, and trained the datasets separately by SVM (Support Vector Machine) model and KNN (k-nearest neighbor) algorithm. The two algorithm models used are described below:

(1) SVM (Support Vector Machine) is a common method of discrimination. In machine learning, it is a supervised learning model that is generally used for pattern recognition, classification, and regression analysis. The SVM maps the sample space to a high-dimensional or even infinite dimensional feature space through a nonlinear mapping, so that the problem of nonlinear separability in the original sample space is transformed into a linearly separable problem in the feature space. In short, it is dimension raising and linearization. Dimension raising is a method of mapping the samples to a high-dimensional space. As for regression and other issues, it is very likely that the sample set cannot be linearly processed in the low-dimensional sample space; while in the high-dimensional feature space, it can be mapped by a linear super-dimension plane to realize linear partitioning. The SVM method cleverly solves the complex computational problem caused by dimension-raising: it applies the expansion theorem of kernel function without knowing the explicit expression of nonlinear mapping; to some extent, it avoids the "dimensionality Disaster "problem.

We used polynomial kernel functions as the kernel functions of the support vector machine in experiments.

(2) KNN is one of the simplest classification methods in data mining classification technology. The so-called k-nearest neighbor indicates that the number of the nearest neighbors is k. That is to say each sample can be represented by its closest k neighbors.

The idea of this method: if a sample has the k most similar (i.e. the nearest neighbor in the feature space) samples in a feature space belonging to a certain category, the sample also belongs to this category. In the KNN algorithm, the selected neighbors are already correctly classified objects. The method determines the category to be sub-sampled according to the category of the nearest one or several samples in the class-decision. The KNN method, although also theoretically dependent on the limit theorem, is only relevant for a very small number of adjacent samples in class decision making. Because the KNN method mainly depends on the limited neighboring samples instead of the discriminant domain, it is more efficient than other methods for the cross-over or overlap of the sample sets to be divided.

In the third step, we trained the preprocessed datasets through deep learning and FNN (feedforward neural network).

FNN (Feedforward Neural Network) is the simplest neural network in which neurons are arranged in layers. Each neuron is connected to the neurons only in the previous layer. It operates by receiving the output of the previous level, and outputting to the next level. There is no feedback between layers, which can be represented by a directed acyclic graph. It is currently the most widely used and the one of the fastest growing artificial neural network.

We used three layers of neurons, the first two layers are activated using the relu function and the last is bisected using the sigmoid function.

ANALYSIS

The accuracy rate of each network after experimental training shows as Table 1.

TABLE 1. Accuracy of different networks

	SVM	KNN	FNN
Accuracy	0.9556	0. 9668	0.9894

The accuracy of SVM is 0.9566, the accuracy of KNN is 0.9668, and the accuracy of FNN is 0.9894. By comparison, the accuracy of the feedforward neural network is higher than the other two machine learning algorithms. The loss is as Fig.1.

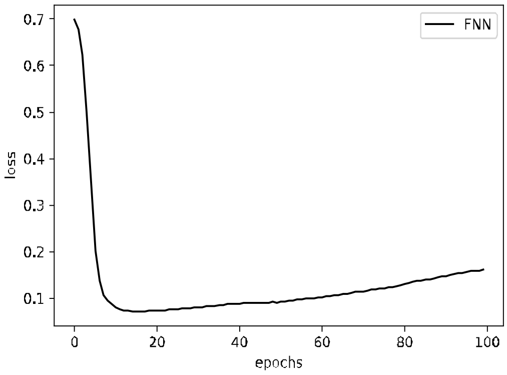

FIGURE 1. The loss of FNN

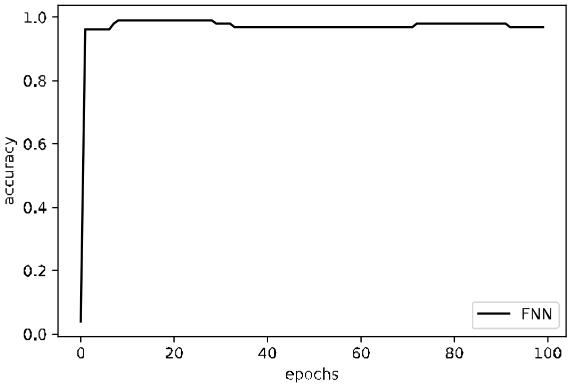

FIGURE 2. Accuracy

The accuracy of FNN and the training results of loss function (cost function) are as Fig.2.

The figure shows that the designed FNN network structure performed smoothly and steadily.

CONCLUSION

In this paper, various algorithms commonly used in the field of machine learning are applied to cervical cancer data classification and comparative research. Based on the extraction of various data features, three types of neural network structures, namely SVM, FNN, and KNN, are used to classify and identify data. The analysis of algorithms are performed in terms of accuracy and results to evaluate. The results show that the three artificial neural networks introduced in this paper have good performance in the auxiliary diagnosis and classification of medical data and provide a basic way to improve the diagnostic intelligence level.

ACKNOWLEDGEMENTS

We are grateful to the following funds:
Student's Platform for Innovation and Entrepreneurship Training Program (1303)
Education Department of Shaanxi Province (15JK1086)
Natural Science Foundation of Shaanxi Province (2015GY009)
Shaanxi University of Science and Technology Dr. Foundation (BJ14-07)

REFERENCES

1. Experimental data: https://www.kaggle.com/loveall/cervical-cancer-risk-classification.
2. Wei Yili, Qiu Yongxiu, Fang Jianling. The study of the diagnostic methods of cervical cancer [J]. clinical and experimental medicine, 2007, 6 (4): 33-35.
3. Schuwei, schuke, PAM, et al, et al. The application of machine learning in early diagnosis and prognosis of cancer [J]. Journal of medical informatics, 2016, 37 (11): 10-14.
4. Cao Wenzhe, Ying Jun, Zhang Yahui, et al. Research on the diagnostic model of prostate cancer based on machine learning algorithm [J]. Chinese medical equipment, 2016, 31 (4): 30-35.
5. Li Rong, Sun Yuan. The application of machine learning in the diagnosis of thoracic cancer [J]. science and technology and engineering, 2011, 11 (20): 4730-4733.
6. Haobin Shi, Xuesi Li, Kao-Shing Hwang, Wei Pan, Genjiu Xu, Decoupled Visual Servoing with Fuzzy Q-Learning, IEEE Transactions on Industrial Informatics, Volume 14; Issue 1, PP: 241-252, 2018(SCI, IF:6.764, area 1, TOP Journal)
7. Haobin Shi*, Zhiqiang Lin, Shuge Zhang, Xuesi Li, Kao-Shing Hwang, An adaptive Decision-making Method with Fuzzy Bayesian Reinforcement Learning for Robot Soccer, Information Sciences, Volume 436-437, pp:268-281, 2018(SCI, IF: 4.832, area 1, TOP Journal)