Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/29/20230997

Signal processing and machine learning in healthcare

An Ping

Sun Yat-sen University (Shenzhen Campus), Shenzhen, Guangdong, China, 518107

pingan3@mail2.sysu.edu.cn

Abstract. Machine learning is the field of artificial intelligence and its important branch. With the continuous innovation of technology, its application in the medical field is increasingly extensive and in-depth. In view of the current human eye discrimination instability and lack of experience, this paper proposes the application of machine learning method, through the training of various attributes of breast cancer data, so that the breast cancer diagnosis system can automatically diagnose malignant breast cancer patients, reduce the influence of human operation on the existence time and experience.

Keywords: machine learning, breast cancer, safety.

1. Introduction

In the Bayesian algorithm, -NN algorithm and SVM algorithm based on machine learning, Bayesian learning is a method to calculate the probability of hypothesis based on the prior probability of hypothesis and the probability of different data observed under a given hypothesis. It can ensure the minimum classification error statistically. The two most basic requirements for using this method for classification are: ① The number of categories of classification problems is known; ② The correlation probability distribution is known or can be solved. However, the Bayesian learning algorithm needs to estimate some probability values in the problem according to the background knowledge and the reference distribution, and the calculation cost is relatively high in some cases. Bayes method is applicable to the following situations: the decision problem can be described in the form of probability, and all probability structures are known.

The -NN algorithm is to establish a classification method without assumptions on the form of function [1]. In general, the -NN rule is used most frequently for problems involving multidimensional features. Since the mail is divided into different categories in this paper, this method can obtain good classification effect.

SVM algorithm is a new machine learning method based on statistical learning [2]. For learning tasks with a limited number of trainings, if the precision of the training set and the capacity of the machine are properly balanced, the optimal generalization ability can be achieved, so that the learning problem can be effectively solved. Moreover, the complexity of SVM algorithm has nothing to do with the sample dimension of the transform space. It's described by the number of support vector machines. However, the classical SVM algorithm only analyzes the problem of binary classification, and in the practical application of data mining, it is generally to solve the problem of multi-value classification, the traditional SVM algorithm has certain limitations, and the accuracy is not very high.

DOI: 10.54254/2755-2721/29/20230997

2. Research on classification algorithm based on machine learning

2.1. Machine learning

Machine learning is a multi-field interdisciplinary subject that has emerged in recent 20 years, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and many other subjects [3]. Machine learning mainly takes artificial intelligence as the research object, especially how to improve the performance of algorithms in experiential learning [4]. A computer program is said to learn from experience E with respect to some class of tasks T and. A computer program is said to learn from experience e with respect to some class of tasks t and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.). Machine learning theory focuses on designing and analyzing algorithms that allow computers to "learn" automatically. The machine learning algorithm is a kind of algorithm that obtains rules from data automatically and uses the rules to predict unknown data.

With the development of computer technology, human beings' ability to collect and store data has been greatly improved, and a large amount of data has been accumulated in various fields of scientific research and social life. How to analyze these data and find the underlying rules has become a common topic in all fields, which makes machine learning technology attract more and more attention [5].

A problem solving system can be defined in the form of triples: where, is the input space of the problem; Is the solution space of the task; Is the solution mechanism of the problem, to some extent, it is a function of. According to Tom M. itchell's point of view, given a certain type of task, related performance and experience, a computer program learns from experience and uses the learning results to improve its performance so as to achieve self-perfection, then the program is said to have learning ability. Therefore, machine learning system is a problem solving system, but machine learning is special, machine learning system can be described in the form of four tuples:, where: training experience, is the input of learning system, can be expressed as, is the input of problem solving system, is its output[6]; : Learning tasks closely related to problem solving tasks, such as general assumptions, parameter control of problem solving system, classification rules and problem solving strategies, etc. : Learning the output of the system, assumptions used by the problem solving executive components; : Evaluation of the performance of a problem solving system, which can be global or local. The goal of a learning system is to improve performance when it is used in a problem solving system. A typical machine learning system architecture is shown in Figure 1-1:

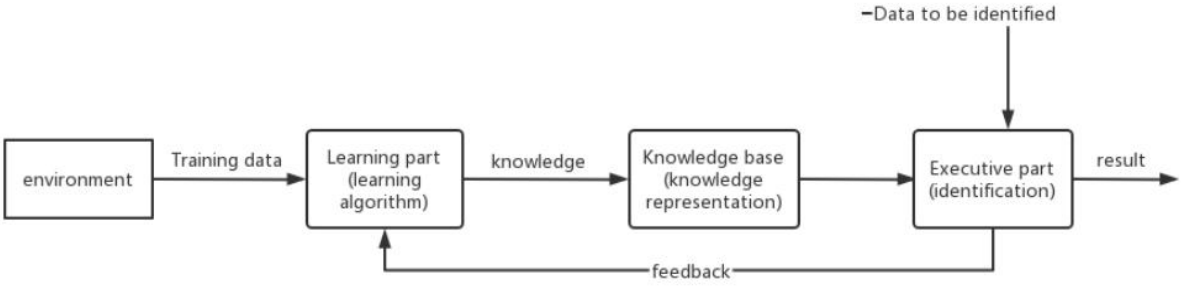

Figure 1-1. Machine learning system architecture.

Figure 1-1 shows the basic structure of a machine learning system. The environment provides some information to the learning part of the system as training data, and the learning part uses the selected learning algorithm to learn the acquired information. The learning process may have an impact on the environment and obtain new training data from the environment. After learning, the acquired knowledge is expressed in different forms according to different learning algorithms, such as rule base, expression, table or neural network, which is collectively referred to as knowledge base. The execution part mainly uses the learning results to process the identification data, obtain the corresponding identification results,

DOI: 10.54254/2755-2721/29/20230997

and feed the execution results back to the learning part for further learning and improvement of system performance [7].

2.2. Classification of machine learning methods

According to the experience contained in the input training data, machine Learning methods can be divided into the following three types: Supervised Leaning, Unsupervised Leaning and Reinforcement Learning. The training data of supervised learning includes the real output of samples, so the real output can be used to evaluate the merits and demerits of the learning structure, such as the disease category of known training samples in medical diagnosis problems and the real price of known training samples in market price prediction problems. There is a lack of empirical knowledge in the training data of unsupervised learning, so it is necessary to establish an evaluation method for the output in the learning process [8]. The empirical knowledge of reinforcement learning comes from the feedback of the environment during the learning process, sometimes with a certain delay.

Supervised learning plays a very important role in machine learning methods, and many classical learning algorithms have been produced, such as Bayesian networks, -NN nearest neighbor algorithms, support vector machines, decision trees and artificial neural networks. This paper introduces several important algorithms in supervised learning.

2.2.1. Bayesian algorithm. The hypothesis and the variables to be solved follow a certain probability distribution, and the probability distribution of the variables and the data observed in the experiment can be reasoned to make the optimal decision, which is Bayesian learning, which provides a probabilistic means and quantitative method for measuring the confidence of multiple hypotheses, and also provides a theoretical framework for the analysis of other algorithms based on probability theory [9].

The Bayesian formula is shown in Formula (1.1). If prior probability, evidence factor and likelihood function are known, then the posterior probability is

$$P(h \mid D) = \frac{P(h \mid D)P(h)}{P(D)} \tag{1.1}$$

Bayesian classifier was proposed by Lewis in 1992. Although it was assumed that the classifier had strong independence, it still achieved good classification effect in the practical application of text classification. Moreover, Lewis and Ringuete et al. proved that the Bayesian classifier had better performance through experiments. Bayes algorithm is not only simple in principle, but also fast in operation and high in classification accuracy, so it has been widely used in the field of text classification and information retrieval.

2.2.2. -NN nearest neighbor method. -NN(-Nearest Neighbor), also known as the Nearest Neighbor method. The principle of this method is to assume that there are many sample data in a classification, and a sample data belongs to a specific classification indicated by the class label. The distance between each sample data and the data to be classified can be calculated through a specific distance calculation formula. And take the sample data closest to the data to be classified. If a certain category occupies the majority in this sample data, the data to be classified belongs to this category. When classifying mails in mail processing, the nearest (most similar) mails in the data set of new mails and training mails are obtained through calculation, and then the category to which the mails belong is determined as the category to which the new mails belong.

In the E-mail classification system, all mail content information is expressed in the form of feature vectors under the vector space model, so the calculation of mail similarity (distance) is transformed into the calculation of mail feature vectors. Generally speaking, the hyperplane of a two-dimensional training set is shown in Figure 1-2.

DOI: 10.54254/2755-2721/29/20230997

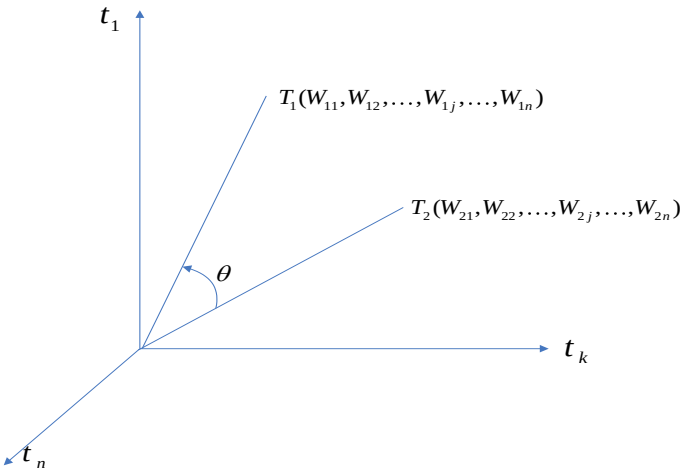

Figure 1-2. Feature representation in vector space.

Where, and respectively represents two different mail feature vectors, represents the weight value of the first feature item of the first mail, and represents the Angle between the two mail feature vectors.

Common methods for calculating vector similarity include Euclidean distance, cosine of included Angle and inner product of vector. Corresponding calculation formulas are shown in Formula (1.2), (1.3) and (1.4).

Euclidean distance:

$$Sim(d_1,d_2)=\sqrt{\sum_{i=1}^{n}(W_{1i}-W_{2i})^2} \tag{1.2}$$

Where, and respectively represent the weights of corresponding feature words in the feature vectors of mail and. The smaller the result calculated by this method, the smaller the distance between two mails and the greater the similarity of mails.

Inner product of vectors:

$$Sim(d_1,d_2)=\sum_{i=1}^{n}(W_{1i}*W_{2i}) \tag{1.3}$$

Where and respectively represent the weights of corresponding feature words in the feature vectors of text and. Since the inner product is the projection of one vector onto another vector, if the inner product is used to calculate the similarity, the greater the inner product, the greater the similarity of the two messages.

Included Angle cosine:

$$Sim(d_1,d_2)=\frac{\sum_{k=1}^{M}W_{1k}\times W_{2k}}{\sqrt{(\sum_{k=1}^{M}W_{1k}^2)(\sum_{k=1}^{M}W_{2k}^2)}} \tag{1.4}$$

Where and respectively represent the weights of corresponding feature words in the feature vectors of mail and. If the Angle between two vectors is smaller, the cosine value is larger, and the mail represented by two vectors is more likely to belong to the same category. Conversely, the smaller the cosine, the less likely it is that the messages represented by both vectors belong to the same class.

DOI: 10.54254/2755-2721/29/20230997

The system designed in this paper is to calculate the similarity of the feature vector between two mails through the included Angle cosine.

The specific steps of -NN algorithm are as follows:

1. Feature words in training emails are described in the form of feature vectors;

2. According to the feature words, Chinese word segmentation and feature extraction are carried out for the content of the new mail, and the feature words in the new mail are described in the form of feature vector;

3. Formula (1.4) is adopted to calculate the mail that is most similar to the new mail in the training mail set. In general, an initial value is set to determine the value, and then the value is constantly adjusted in the experimental test.

4. According to the mail in Step 3, calculate the weight of each category in turn and use formula (1.5) to calculate:

$$p(\bar{x}, C_j) = \sum_{\bar{d}_i \in KNN} Sim(\bar{x}, \bar{d}_i) y(\bar{d}_i, C_j) \tag{1.5}$$

Where, the similarity between the feature vector of the new mail and is the same as Step 3, category attribute function, with the value of 0 or 1.

5. Compare the weights calculated in Step 4. The category with a larger weight is the category of the new mail.

2.2.3. SVM support vector machine algorithm. Support Vector Machines (SVM) were proposed by Vapnik and collaborators in 1992 and have since gained widespread attention in machine learning and have become one of the standard tools in machine learning and data mining. The theory of support vector machine is based on the VC dimension theory of statistical learning theory and the principle of minimum structural risk. It seeks the best compromise between the complexity of the model and the learning ability according to the limited sample information and expects to obtain the best generalization ability [2]. The linear classification problem of the two types of problems of support vector machine is actually the optimal classification surface problem under the condition of linear separability, then there is the following mathematical model:

If there is a real-valued function, its operation mode is, when, it will be divided into positive class, otherwise divided into negative class. When is a linear function, it can be written as formula (1.6):

$$\begin{aligned} f(x) &= \langle w \cdot x \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned} \tag{1.6}$$

Its geometric interpretation is that the hyperplane is defined by, and the input space is divided into positive class and negative class. The positive class includes the points above the hyperplane, while the negative class includes the points below the hyperplane, and the hyperplane constantly shifts with changes. Generally speaking, the hyperplane of two-dimensional training set is shown in Figure 1-3.

DOI: 10.54254/2755-2721/29/20230997

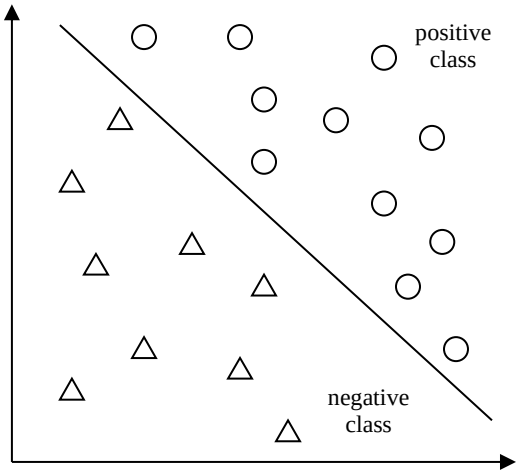

Figure 1-3. Hyperplane of two-dimensional training set.

Sometimes it is difficult for the original input space to be linearly separable, thus increasing the difficulty of the learning task. Therefore, the input space is generally transformed by mapping to make it linearly separable in high dimensions. Because of its strong theoretical basis, support vector machine algorithm has achieved good results in the field of text classification.

2.2.4. Enhance learning. Reinforcement learning is characterized by trial-and-error interactions with the environment to determine and optimize the choice of actions to achieve the so-called sequential decision task. In this task, the learning mechanism interacts with the environment by selecting and executing actions that lead to a change in system state and possibly some kind of reinforcement signal (immediate reward). Reinforcement signals are quantified rewards and punishments for system behavior. The goal of systematic learning is to find an appropriate action selection strategy, that is, to choose which kind of action in any given state, so that the generated action sequence can obtain some optimal result (such as maximum cumulative immediate return).

In comprehensive classification, experiential inductive learning, genetic algorithm, association learning and reinforcement learning belong to inductive learning. Experiential inductive learning uses symbolic representation, while genetic algorithm, association learning and reinforcement learning uses subsymbolic representation. Analytical learning belongs to deductive learning.

In fact, analogy strategy can be regarded as a synthesis of inductive and deductive strategies. Therefore, the most basic learning strategies are only induction and deduction.

From the perspective of learning content, learning by induction strategy is the induction of input, so the knowledge learned is obviously beyond the scope of the original system knowledge base, and the learned result changes the knowledge deductive closure of the system, so this type of learning can also be called knowledge level learning. Although the knowledge learned by deductive strategy can improve the efficiency of the system, it can still be contained by the knowledge base of the original system, that is, the knowledge learned cannot change the deductive closure of the system. Therefore, this type of learning is also called symbolic level learning.

3. Automatic diagnosis of diseases using machine learning

No matter text data in medical data or image data such as CT and X-ray, the method of using machine learning to diagnose diseases is essentially a problem of using the automatic learning ability of machine learning to realize supervised classification. From the research in Chapter 1, we know that machine learning has been deeply applied to the field of intelligent diagnosis. The diagnostic model learned not only has high efficiency but also achieved considerable ability in accuracy, and the overall robustness and generalization ability have been enhanced. At present, the machine learning methods applied to

DOI: 10.54254/2755-2721/29/20230997

disease intelligent diagnosis include linear regression model, support vector machine, artificial neural network, decision tree, Bayesian classification, integrated learning, K-nearest neighbor classification algorithm, and deep reinforcement learning. At the same time, the improved models based on these models also get better prediction results on some diagnostic problems. Different classification models have different characteristics, and their performance in different scenarios and different data sets is not satisfactory. Different algorithms should be selected according to specific tasks to effectively solve the problem. In general, the machine learning diagnosis process is consistent, as shown in Figure 2-1.

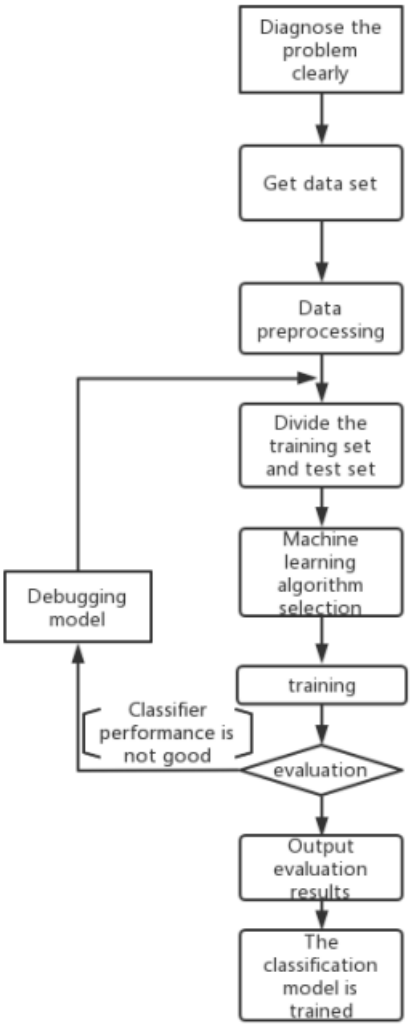

Figure 2-1. Machine learning diagnosis process.

As can be seen from Figure 2-1, the most important steps of disease diagnosis are:

1. Get the data set. Using machine learning to diagnose diseases, the most important thing is the support of big data.

Different classification tasks, especially in medical disease diagnosis, require much larger and richer data sets. In addition, each feature of the data also has an important impact on the overall classification task. On the one hand, it can collect data artificially and master important features, which can promote the later intelligent diagnosis task to achieve good classification effect. On the other hand, all kinds of open source data sets are widely used at present, such as UCI data set, GoogleTrends data set, Kaggle data set, Imagenet, MNIST, Reddit /r/ data set, etc. The UCI is one of the most famous open-source datasets in the world; Imagenet is one of the most famous databases in the image category. Many

DOI: 10.54254/2755-2721/29/20230997

researchers have verified various algorithms related to image processing on Imagenet, which promotes the development of deep learning.

There are also a large number of datasets for a specific field, such as MIT face recognition, image separation, song databases, geological data, image processing COCO, video datasets, government data, and so on.

2. Data preprocessing. Before selecting the data set for training, it is necessary to check whether there is any missing sample data and whether the number of samples is even, etc., and then processing the data input to the model to make the data fit for the model for training.

Generally speaking, the methods of preprocessing vary with data. Commonly used are processing missing data, normalization of numerical data, de-averaging, dimensionality reduction, and so on. These are described in detail in the next section.

3. Separate test set and training set. After the sample data is preprocessed, it is divided into two sample subsets in a certain proportion. The training set is used to train the suitable model, and the test set is used to evaluate the classification ability. The problem of disease diagnosis can be summarized as supervised machine learning. Usually, the sample number of training set is much larger than that of test set.

4. Machine learning algorithm selection. Since disease diagnosis is also a classification problem, the classification algorithms commonly used in automatic diagnosis include SVM, decision tree, artificial neural network, Bayesian classification, integrated learning, K-nearest neighbor classification algorithm, deep reinforcement learning, etc., and various improved algorithms on the basis of these simple algorithms. At present, Transfer Learning, enhancement learning and various improved deep learning algorithms have been widely applied, and the application effect is very good in disease diagnosis.

5. Train. The essential problem of training various classification models is to find the ideal parameters to make the model's prediction more accurate. In general, the ideal parameter combination is selected by the loss function in the mapping relationship between eigenvalue x and weight y. This process is implemented based on a training set.

6. Evaluate. For a good algorithm model, reasonable and effective evaluation index is the key. For each classification task, only a good evaluation index can provide valuable guidance for the selection of algorithm model and later model debugging. Otherwise, an inappropriate evaluation method will only waste time. Accuracy is the most intuitive and useful evaluation index, while confusion matrix, accuracy rate, recall rate and other indicators. This process is implemented based on test sets.

7. Debug the model. Judging by the evaluation index if there is error in the model, it is necessary to debug the model and the training process. The identification of overfitting problem or underfitting problem is generally carried out through the debugging process of the number of sample training sets, the number of characteristic samples, the selection of characteristic values, the adjustment of parameters or the thinking of cross verification.

4. Analysis of breast cancer diagnosis process and results based on machine learning

Machine learning is about making machines replace or assist humans to do a better job. Breast cancer data has typical classification properties. By constructing a high-performance classification model, it can determine whether the person coming for diagnosis has a malignant tumor. There are various models for classification, but the basic steps are as follows:

(1) Analysis of data characteristics:

Because of the variety of data types, there are a variety of machine learning models. Different data types suit different machine learning models, and the results achieved through machine learning will be different. This breast cancer data set has clear attributes and corresponds to confirmed diagnosis results, so it can be seen that this data set is more suitable for classification algorithm training.

(2) Data preprocessing:

We used the data 70% as the training set and 30% as the test set. The isnull ().sum (axis=0) method in sk-learn library of python was used to calculate missing values, and then the dropna (inplace=True)

DOI: 10.54254/2755-2721/29/20230997

method was used to remove missing values. Finally, 683 complete valid data were selected from 696 data.

(3) Construction of classification model:

In logistic regression algorithm, the threshold for judging probability is generally 0.5, but we can set the threshold to 0.3 to improve the "sensitivity" of the model.

In terms of decision tree algorithm parameters, splitter='best' is used to select the optimal segmentation feature and segmentation point. By traversing the depth of the tree, the optimal depth of the tree is selected as 3.

Support vector machine. Since the support vector machine algorithm is time-consuming, we use logistic regression modeling and output the coefficient to analyze the importance. It was found that the absolute value of the final output value of the attribute boring chromosome was the largest. Therefore, in order to save time, only the attribute sample of the attribute boring chromosome was selected. In this experiment, Linear kernel function was selected as the modeling model. Through repeated experiments, it was found that penalty coefficient C=100 of the objective function was optimal.

The most important thing for KNN algorithm is to select an appropriate parameter K, which is selected in this experiment. coss_ val_ score method, cross validation is used to evaluate the predictive performance of the model, especially the performance of the trained model on the new data, which can reduce the overfitting to some extent. Add loops through which parameters are constantly changed and cross-validation is used to evaluate the capabilities of different parameter models. Finally, the model with the best selection ability has the best performance when K is 5.

(4) Performance evaluation: According to the evaluation indexes of the above model, there are mainly seven: confusion matrix, accuracy rate, accuracy rate, recall rate, the harmonic average of accuracy rate and recall rate of F1-SCORE), AUC (area under ROC curve), and time consuming (modeling time).

The following table compares the classification results based on breast cancer data:

AUC (Area Under Curve) is defined as the product under ROC curve. AUC expresses probability, so the value of the area under curve will not be greater than 1. Since the ROC curve is generally below y=x, the value interval of AUC can be calculated as [0.5]. The ROC curve is difficult to tell which classifier is more effective, but the AUC is a number, so it is very clear which classifier is more effective.

It can be seen from the results of the above classification samples that KNN is the best classification model for breast cancer data, followed by logistic regression, SVM, and decision tree model. The misjudgment rate of the four models can be calculated by using the classified data in the confusion matrix, among which KNN is 0.034, decision tree is 0.049, SVM is 0.049, and logistic regression is 0.044. From the perspective of modeling time, decision tree and KNN take the shortest time, followed by SVM and logistic regression. Although both decision tree and KNN consume the shortest time, the accuracy of KNN is 2% higher than that of decision tree.

(5) Model application: The results show that among the four machine learning models, the KNN algorithm model has the lowest misjudgment rate and the highest accuracy in breast cancer classification. The AUC data obtained by the ten-fold cross-validation method shows that the KNN model has the best performance. Finally, KNN classification model with low misjudgment rate, high accuracy rate and short time is selected.

5. Conclusion

Machine learning, as an emerging research hotspot in the current society, has been widely used in various industries. On the basis of understanding the basic situation of machine learning, this paper studies the machine intelligence technology from the standpoint of machine learning technology, discusses the application of machine learning in various industries, and promotes the further application of machine learning in the medical industry.

References

[1] Zhao J Y. Research and Application of machine learning methods for Health evaluation [D].

DOI: 10.54254/2755-2721/29/20230997

University of Electronic Science and Technology of China,2016.
[2] LIU Qiaoli. Research on Automatic Disease Diagnosis Based on Machine Learning [D]. Yantai University,2019.
[3] Liu Shuo. Breast cancer data analysis and computational modeling [D]. Fujian Normal University,2018.
[4] Zheng Y W. Breast cancer diagnosis based on feature selection and support vector machine [D]. Taiyuan University of Technology,2019.
[5] Wang Xin. Research on Cancer Detection Based on Machine Learning [D]. Lanzhou University,2019.
[6] Kong DF. application of machine learning in breast cancer diagnosis [J]. Information Communication,2019(07):18-21. (in Chinese)
[7] Zhang S J. Study on the method of breast cancer detection based on mammogram image [D]. Beijing Jiaotong University,2013.
[8] Qin W J. Research on key issues of medical image segmentation based on machine learning and its application in tumor diagnosis and treatment [D]. University of Chinese Academy of Sciences (Shenzhen Institute of Advanced Technology, CAS),2019.
[9] Huo S H. Breast tumor recognition based on machine learning [D]. North University of China,2017. (in Chinese)