Scholars Academic Journal of Biosciences
Abbreviated Key Title: Sch Acad J Biosci
ISSN 2347-9515 (Print) | ISSN 2321-6883 (Online)
Journal homepage: https://saspublishers.com/journal/sajb/home

OPEN ACCESS

Mathematics & Physics

Prediction of SNP Pathogenic Site Based on Naïve Bayes Method

Xue Wu, Baoguang Tian*

School of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

***Corresponding author:** Baoguang Tian

| **Received:** 01.03.2019 | **Accepted:** 08.03.2019 | **Published:** 14.03.2019
DOI: 10.36347/sajb.2019.v07i03.001

Abstract

Original Research Article

SNP site is an important basic variation data, which has the characteristics of large amount of data and uniform distribution. It is widely used in complex disease research, and the data mining of SNP pathogenic site by machine learning method has become research focus in field of bioinformatics. In this paper, we present a new SNP pathogenic site prediction method based on the naïve Bayes. First, we select 1000 samples of all the SNP sites (9445 sites) information on a chromosome fragment, and the base (A, T, C, G) of each SNP site has three manifestations, which are converted into 0,1,2 numerical codes. Secondly, 447 possible SNP pathogenic sites and one abnormal SNP site are selected by chi-square test according to the encoded information and information of those samples with genetic disease. Finally, the naïve Bayes model is established on 1000 samples to predict SNP pathogenic site. Five-fold cross validation indicates our method achieves superior performance with an ACC of 84.64% and MCC of 0.6937, respectively. Compared with those of other machine learning methods, the results show that the prediction performance of naïve Bayes model is better than that of K-nearest neighbor (KNN), AdaBoost, support vector machine (SVM) and random forest (RF) model.

Keywords: SNP pathogenic site prediction, chi-square test, naïve Bayes, support vector machine, AdaBoost, K-nearest neighbor.

INTRODUCTION

Single nucleotide polymorphism (SNP) refers to the polymorphism that caused by the variation of a single nucleotide (A, T, C, G) in the genomic DNA sequence, which is the most widely distributed and rich genetic information polymorphism in the human genome [1,2]. A large number of existing SNP site determines the interpersonal personality differences, and there is a great correlation with the complex diseases. Nowadays, SNP site research is generally considered to be an important step in the application of the human genome project [3]. Thus, by comparing the health status of the sample with the site coding information, locating the SNP site associated with the disease in the chromosome or gene position which can help the researcher understand the genetic mechanism of the trait and some disease, this approach prevents some genetic diseases from occurring by interfering with SNP pathogenic sites [2]. However, the SNP site data is high-dimension and the number of samples is limited, so that to detect disease-related SNP site in the human body research work progress is slow. With the rapid development of bioanalytical techniques and computational techniques, the machine learning method which is used to predict SNP pathogenesis site has become hotspot in the field of bioinformatics [4,5].

In the present research, the researchers have used a variety of machine learning methods in SNP pathogenic site prediction and identification, including the K-nearest neighbor (KNN) [6], neural network [7,8],decision tree (DT) [7,8], support vector machine (SVM) [9] and other supervised learning methods [10,11], which are characterized by integration site characteristics, automatic classification. In the data mining prediction of SNP site information also has the application of unsupervised learning. Lee *et al*. [12] proposed a tag SNP site selection method based on the Bayesian network, which did not depend on the haplotype sequence structure of the adaptive multi-allele tag SNP selection method, the network model selected the tag SNP site from the input SNP site sequence information data and outputs the haplotype sequence of all SNP sites. Patil [13] proposed the method of based on the haplotype block selection pathogenic site. It is based on the number of human haplotype less than the theoretical number of the basic principles, the genome sequence data was divided into multiple discrete monomer blocks. In these blocks, it is possible to identify the set of minimum SNP sites that can distinguish all (or most) haplotypes in each haplotype as the pathogenic SNP set. Mahdevar *et al*. [14] proposed a heuristic method, based on Genetic Algorithm (GA). The SNPs were input and the binary vectors of length n were used to represent the individuals in the GA. The fitness function was based on the minimum number of tag SNPs and combined the Shannon entropy function to design two new crossovers and mutations, output a set of tag SNPs and used them in various simulation and experimental data to find the most likely tag SNP within an acceptable time. Phuong *et al*. [15] used the SNP site as a feature, and defined the range distance according to the LD association between the sites, and then obtained the tag SNP site by the feature selection method. Halldorsson *et al*. [16] proposed information SNP site selection method based on accuracy prediction. The information SNP was defined as a set of SNP sites that accurately reconstruct the remaining SNP sites or information site, selected the SNP site set of information to maximize the prediction of the remaining tagged SNP, reconstructed the corresponding haplotype sequence, and evaluated the information SNP site prediction capability.

This paper presents a new method of SNP pathogenic site prediction based on naïve Bayes. Firstly, the base (A, T, C, G) of 9445 SNP sites three manifestations are converted into 0, 1, 2 numerical codes. Secondly, the 1000 samples divide into 500 healthy samples and 500 unhealthy samples. The frequency of “0”, “1” and “2” at each site is summarized, calculating the p-value of 9445 SNP sites by the chi-square test. In the case of a significance level of 0.05, it selects 447 possible SNP pathogenic sites and 1 abnormal SNP site, lowering the 9445 SNP sites to 448 SNP pathogenic sites. Finally, the naïve Bayes model is built up on the 1000 sample data using five-fold cross validation to the predict SNP pathogenic site. The experimental results show that our method achieves superior performance with an ACC of 84.64% and MCC of 0.6937, respectively. Compared with the main results of other machine learning methods, the prediction accuracy of naïve Bayes is higher than that of KNN, AdaBoost, SVM and RF method.

MATERIALS AND METHODS

Datasets

This dataset includes phenotypic information about 1000 samples of genetic disease and coding information for 9445 SNP sites. The sample information contains 500 people who are not suffering from genetic disease and 500 people with genetic disease, with “0” for healthy people, “1” for unhealthy people. The site information contains the coding information of 9445 SNP sites under the 1000 sample. The SNP site name begins with rs and uses the encoding of the base (A, T, C, G) to represent the information of each sites, such as the SNP site rs3094315, different samples of the code are T and C combinations. There are three different encoding methods TT, TC and CC. Similarly, although the combination of other SNP sites of the base is different, but there are only three different coding. These datasets containing genetic disease information can be downloaded from the web site http://gmcm.seu.edu.cn/31/list.htm.

Feature extraction

Non-coding SNP site information can be seen as a special string, this paper studies the data set to provide phenotypic information and 9445 SNP site coding information of 1000 samples, through the sample health and site coding comparative analysis to study SNP pathogenic site to find the genetic mechanism of disease or traits, compared the classification performance of different machine learning methods.

At each SNP site location; there are three different coding schemes for different samples. In order to facilitate the data analysis, we select the frequency of occurrence of characters as a feature, with the value of 0, 1, 2 pairs of non-coding SNP site information to replace. Firstly, we calculate the frequency of the three coding modes of 1000 samples at 9445 SNP sites and then substitute 1 for the heterozygous, 0 and 2 instead of the higher frequency and lower frequency homozygous, 9445 sequence features as raw data, as shown in equation (1).

$$SNP_i = \begin{cases} 0 & \text{high frequency of homozygous} \\ 1 & \text{heterozygote} \\ 2 & \text{low frequency of homozygous} \end{cases} \quad (1)$$

Table-1: Partial information before and after numerical coding

Before				After			
rs3094315	rs3131972	……	rs7545865	rs3094315	rs3131972	……	rs7545865
TT	CT	……	GA	0	1	……	1
TC	CT	……	GG	1	1	……	0
TT	TT	……	GA	0	2	……	1
TT	CC	……	GG	0	0	……	0
TC	CT	……	GA	1	1	……	1
……	……	……	……	……	……	……	……
TC	CT	……	AA	1	1	……	2

In Table 1, the first line is the name of 9445 SNP sites, which are all beginning with rs. The remaining four rows of the remaining rows of 1000 samples of part of the text encoding information, the left side of the remaining four columns for the 1000 samples part of the text encoding information, each SNP site is up to three encoding combination, the right side of the part of the encoded numerical information, encoding 0,1,2. Such as the SNP site rs3094315, there are only three text encoding (TT, TC, CC), since "TT" appears at a frequency higher than the frequency of "CC", replace "TT" with "0", replace "CC" with "2", replace "1" with heterozygote "TC", and the remaining SNPs are similar to the sites.

Chi-square test

The chi-square test [17] is a hypothesis test method for counting data. It essentially compares two or more frequencies to detect the difference between the actual frequency at a certain significance level and the expected frequency based on a theoretical model or distribution feature hypothesis.

Because the differences in the sample are caused by the intrinsic factor rather than the sampling error, the chi-square value is large, and it's the corresponding p-value that the probability of reflecting the differences in the sample caused by the sampling error is small. The differences of sample are "significant" or "highly significant". On the contrary, the smaller the chi-square value, the greater the p-value, said the two samples of the difference "no significant".

In practice, preparatory calculations can be made using the contingency table model. Suppose there are two categorical variables X and Y, and their ranges are (x_1, x_2) and (y_1, y_2, y_3), where the sample frequency is shown in Table 2.

Table-2: Chi-square test six grid table

	y_1	y_2	y_3	Total
x_1	a	b	c	$a+b+c$
x_2	d	e	f	$d+e+f$
Total	$a+d$	$b+e$	$c+f$	$a+b+c+d+e+f$

The chi-square values of χ^2 are calculated from the data in Table 2:

$$\chi^2 = n(\frac{a^2}{(a+b+c)(a+d)} + \frac{b^2}{(a+b+c)(b+e)} + \cdots + \frac{f^2}{(d+e+f)(c+f)} - 1) \qquad (2)$$

where the total number of samples is $n = a+b+c+d+e+f$.

After calculating the chi-square value of χ^2, the chi-square value table is checked by the degree of freedom $n' = (2-1)(3-1)$ to determine the significance of the difference between the groups. When $X^2 > X^2_{0.01}$, $P < 0.01$, the differences are highly significant; when $X^2_{0.01} > X^2 > X^2_{0.05}$, $0.01 \le P \le 0.05$, the differences are significant; when $X^2 < X^2_{0.05}$, $P > 0.05$, the differences are not significant.

Machine Learning Method

Naïve Bayes

Naïve Bayes (NB) is a typical statistical method used to predict the probability that a sample belongs to a particular class [18]. The naïve Bayes classifier is based on the Bayesian statistics and decision theory. It can effectively deal with incomplete or partial data loss dataset, which is an ideal expression pattern of combining a priori knowledge and data. The advantage is that the model can be explained with high accuracy. The naïve Bayes model is based on the Bayesian theorem, which reduces the computational cost through the assumption of conditional independence and predicts that the unknown data samples belong to the highest posterior probability class. Some studies have shown that the conditions of the independence assumption of the naïve Bayes classification model can't be satisfied, but the classification performance in some fields can still be comparable to the classical algorithm of decision tree algorithm and KNN. At present, the naïve Bayes method has been widely used in medical diagnosis, bioinformatics, financial analysis and other aspects [19].

The definition of the naïve Bayes classification is as follows:

Step1: Let $x = \{a_1, a_2, \ldots, a_m\}$ is a feature to be classified, and each a_i is a characteristic attribute of x .

Step2: There is a category set $C = \{y_1, y_2, \ldots, y_n\}$.

Step3: To calculate $P(y_1 | x), P(y_2 | x), \ldots, P(y_n | x)$.

Step4: If $P(y_k \mid x) = \max\{P(y_1 \mid x), P(y_2 \mid x), \ldots, P(y_n \mid x)\}$, then $x \in y_k$.

Then the key is how to calculate the probability of each condition in step 3, we can do so:

1) To find a collection of known categories to be classified items, called training sample set.

2) To estimate the conditional probability estimates for each characteristic attribute under each category, which is

$$P(a_1 \mid y_1), P(a_2 \mid y_1), \ldots, P(a_m \mid y_1); \ldots; P(a_1 \mid y_n), P(a_2 \mid y_n), \ldots P(a_m \mid y_n) \qquad (3)$$

3) If each characteristic attribute is conditional independent, then the Bayesian theorem has the following derivation

$$P(y_i \mid x) = \frac{P(x \mid y_i)P(y_i)}{P(x)} \qquad (4)$$

Because the denominator is constant for all categories, it is only necessary to maximize the numerator for comparison. The characteristic attributes are conditional independent, so there are:

$$P(x|y_i)P(y_i) = P(a_1|y_i)P(a_2|y_2)\cdots P(a_m|y_i)P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \qquad (5)$$

The structure of the naïve Bayes method is shown in Fig. 1.

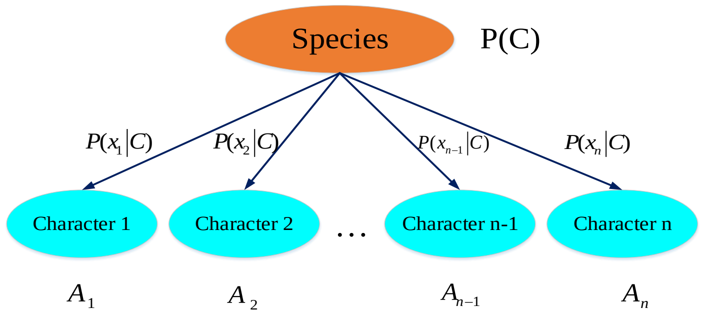

Fig-1: Naïve Bayes classification model structure

In Fig. 1, the leaf nodes $A_1, A_2, \ldots, A_n$ represent attribute variables, and the root node represents the category variable. The principle of classification by the naïve Bayes method, according to the prior probability of an object, the posterior probability is calculated by using Bayesian formula. That is to say, the probability that the object belongs to a certain class, and select the category with the maximum posterior probability as the class to which the object belongs.

AdaBoost algorithm

AdaBoost is an iterative algorithm [20], the basic idea is that weak classifiers are clustered together for the same training set, thus forming a stronger final classifier.

The basic principles of the AdaBoost classification are described below:

Step1: To give a set of training sets for training: $\{x_1, y_1\}, \{x_2, y_2\}, \cdots \{x_n, y_n\}$, where x_i is the input training set, y_i is the result of the classification "0" and "1", "0" for "health", "1" means "unhealth".

Step2: The number of times of the iteration is specified, which determines the number of weak classifiers that are selected to form the strong classifier, and iterates a weak classifier with better sorting ability.

Step3: Initialize the weight of the sample, $w^1 = \{w_{1,1}, \cdots, w_{1,n}\}$, $\omega_{1,i} = d(i)$, where $d(i)$ is the probability of distribution of the initial sample.

Step4: For $t = 1 : T$

1) Initialize the weight, $w^t = w_{t,1}, \cdots, w_{t,n}$.

2) Train weak classifier, the initial sample learning algorithm with the training to learn, get a weak classifier, $h_i = X \to [0,1]$.

3) Calculate the error rate of each weak classifier under the currently determined weight, as shown in equation (6):

$$\varepsilon_i = \sum_{i}^{n} \omega_{t,i} \left| h_i(X_i) - y_i \right| \qquad (6)$$

From the obtained weak classifier to select the smallest error rate classifier, add them to the strong classifier.

4) Select the appropriate weight: $\omega_{t+1,i} = \omega_{t,i}\beta_t^{1-|h_i(x_i)-y_i|}$, if the $i-th$ sample classification is correct, then $\varepsilon_i = |h_i(x_i)-y_i| = 0$, otherwise $\varepsilon_i = |h_i(x_i)-y_i| = 1$, $\varepsilon_i = |h_i(x_i)-y_i| = 1$and $\alpha_t = \varepsilon_t / 1 - \varepsilon_t$.

5) After T cycles of the classification process, we can get T weak classifiers, and then in accordance with the updated weight of the superposition, the final strong classifier is:

$$H(x) = \begin{cases} 1 & \sum_{t=1}^{T} \alpha_i h_i \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t, \alpha_t = \log(\frac{1}{\beta_t}) \\ 0 & other \end{cases} \qquad (7)$$

Support vector machine

Support vector machine (SVM) is a machine learning method based on statistical learning theory and structural risk minimization [22]. SVM has better adaptability and higher classification accuracy in solving the bioinformatics data [11,23,24] that has a small sample, non-linear and high-dimensional features.

SVM is a statistical learning method, which is aimed to search the best classification by the secondary planning and based on the theory of non-linear mapping. It maps the input sample set to the high-dimensional space and constructs the optimal hyperplane. Learning algorithm, and automatically find out the classification of those who have a better classification of support vector, making the super-plane to the distance between the two sets of the largest sample.

Assuming that a sample set $(x_i, y_i), i = 1,...n, x \in R^d, y \in \{+1,-1\}$ is given

$$y_i\left[\left(\omega^t x_i\right) + b\right] - 1 \geq 0, i = 1,...,n \qquad (8)$$

where the classification surface equation is $\omega^t x_i + b = 0$, at this moment, the classification interval is $\rho = 2/\|\omega\|^2$, the ultimate goal is to find a classification surface, so that the two categories of samples can be correctly classified at the same time to ensure that the largest classification interval, here is the smallest $\|\omega\|^2$.

The SVM problem is transferred into the following form:

$$\begin{cases} \min\limits_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l} \varepsilon_i \\ y_i\left(\omega \cdot x + b\right) \geq 1, i = 1,2,...,l \end{cases} \qquad (9)$$

In the case of solving the optimal classification surface in the high dimension space, the inner product operation of the high dimension space can be transformed into the function of the low dimension space by using the appropriate kernel function $k\left(x_i, x_j\right)$, and the corresponding decision function of the optimal classifier is also transformed into:

$$f(x) = \operatorname{sgn}\{(\omega, x) + b\} = \operatorname{sgn}\left[\sum_{x_i \in sv} \alpha_i y_i k\left(x_i, x_j\right) + b\right] \qquad (10)$$

$k\left(x_i, x_j\right)$ is the kernel function, the introduction of the kernel function solves the high dimension problem, which transforms the inner product operation of the high dimension space into the function of the low dimension space.

K-nearest neighbor algorithm

K-nearest neighbor (KNN) algorithm, which is a classical statistical pattern recognition method, one of the simplest machine learning algorithms [25]. The idea is to calculate the similarity between each sample of the test sample and the training sample set, to find the k most similar samples. This “majority voting method” is generally weighted, and the closer the point, the greater the voting weight. A variety of softwares have a variety of weight functions to be selected.

The KNN algorithm’s steps are as follows

- The sample set to be classified is input, the training sample set is re-described according to the feature item.
- For a test sample, the set of test samples is formed according to the characteristic words.
- Calculate the similarity between the test set and the training set, use the angle cosine to represent the distance, the formula is as follows:

$$Sim(d_i, d_j) = \frac{\sum_{k-1}^{M} W_{ik} \cdot W_{jk}}{\sqrt{\sum_{k-1}^{M} W_{ik}^2}\sqrt{\sum_{k-1}^{M} W_{jk}^2}} \qquad (11)$$

4) For the k neighbors of the resulting new sample, we calculate the weights of each class in the new sample in turn, and the formula is as follows:

$$p(\bar{x}, C_j) = \sum_{\bar{d}_i \in KNN} Sim(\bar{x}, \bar{d}_i) y(\bar{d}_i, C_j) \qquad (12)$$

where $\bar{x}$ is the characteristic variable of the new sample, $Sim(\bar{x}, \bar{d}_i)$ is the similarity calculation formula, $y(\bar{d}_i, C_j)$ is the class attribute function, $\bar{d}_i$ indicates that if the class belongs to the positive class, then the C_j function value is 1, otherwise the C_j function value is 0.

5) Compare the weights of the positive classes with negative classes, and divide the samples to be classified into the category with the largest weight.

Random forest method

Random forest (RF) method [26] is a new machine learning algorithm proposed by Leo Breiman in 2001. The idea is to use the bootstrap sample to obtain *k* sample-data sets, train *k* classifiers, and then multiply the results of multiple decision trees. The RF method has good classification performance and high classification accuracy, which is suitable for all kinds of data sets, and has good robustness to feature selection. It is gradually applied to data mining, bioinformatics and the other fields [27,28].

RF is a classifier composed of multiple decision trees $\{h(x, \theta k)\}$, where $\{\theta k\}$ is a random vector of independent and identical distributions. When the sample X to be classified is input, the classification result is determined by the result of the output of a single decision tree. The decision tree determines the final class tag of the input vector X. RF is a nonparametric classification method driven by data that does not require a priori knowledge of classification, simply by classifying rules for learning training for a given sample [29].

RF increases differences of classification models by constructing different training sets, so as to improve the extrapolation ability of the combined classification model. Through *k* wheel training, we get a classification model sequence $\{h_1(x), h_2(x), \cdots, h_k(x)\}$, and then construct them a multi-class model system, using a simple majority vote method. The final classification decision:

$$H(x) = \arg \max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \qquad (13)$$

Where $H(x)$ is the combined classification model, h_i is a single decision tree classification model, and Y is the output variable, indicating that the majority of voting decisions are used to determine the final classification. RF integrates a number of decision trees, needing to set the number of the constructed trees as ntree. Each node is randomly selected and the number of attributes is mtry.

Performance measures

In statistical learning theory, such methods as jackknife test, self-consistency test, independent test, and k-fold cross validation are often used to evaluate the prediction performance [30]. In our study, five-fold cross validation test method are used to examine the performance of the prediction model. Sensitivity (SE), Specificity (SP) and Matthew's correlation coefficient (MCC) and accuracy (ACC) are used to evaluate the results of the prediction system. These measures can be defined as follows:

1) Accuracy (ACC)

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad (14)$$

2) Sensitivity (SE)

$$SE = \frac{TP}{(TP + FN)} \qquad (15)$$

3) Specificity (SP)

$$SP = \frac{TN}{(TN + FP)} \qquad (16)$$

4) Matthew's correlation coefficient (MCC)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \qquad (17)$$

TP, TN, FP and FN in the above evaluation indicators represent true positive, true negative, false positive and false negative. Where TP is the number of positive samples to be correctly predicted, TN is the number of counter-correct samples correctly predicted, FP means that the counter-count is incorrectly predicted as the number of positive samples, and FN indicates that the number of positive samples is incorrectly predicted as the number of anti-samples.

This paper presents SNP pathogenic site prediction method based on naïve Bayes. The calculation flow chart is shown in Fig. 2. Simulation of the experimental environment: Intel (R) Core (TM) i5-4210U CPU @ 1.70GHz 2.40 GHz 4.00GB of memory, with programming RStudio-1.0.136 implementation.

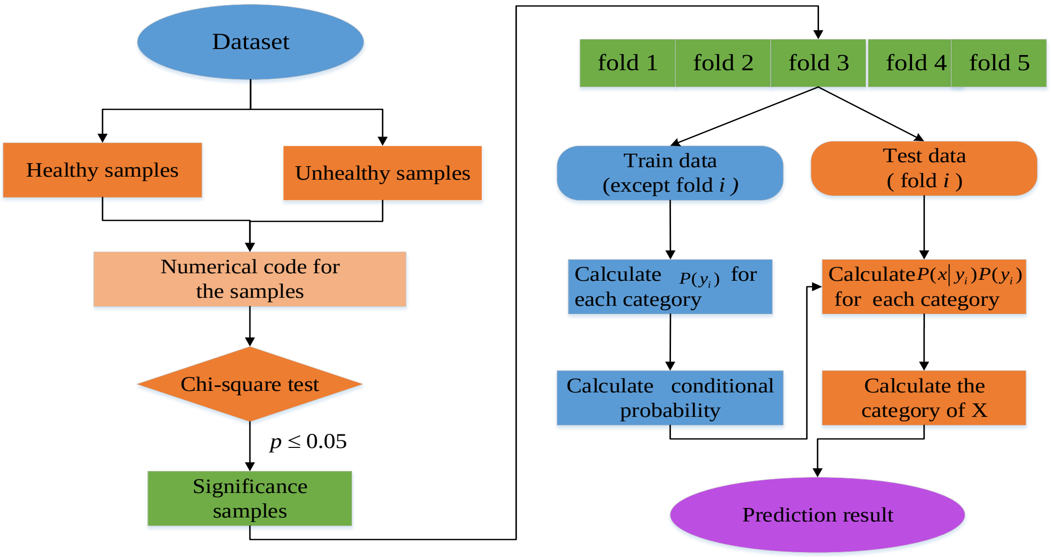
Dataset
Healthy samples
Unhealthy samples
Numerical code for the samples
Chi-square test
$p \leq 0.05$
Significance samples
fold 1
fold 2
fold 3
fold 4
fold 5
Train data (except fold *i*)
Test data (fold *i*)
Calculate $P(y_i)$ for each category
Calculate $P(x|y_i)P(y_i)$ for each category
Calculate conditional probability
Calculate the category of X
Prediction result

Fig-2: Flowchart of SNP pathogenic site prediction based on naïve Bayes method.

The specific steps of the naïve Bayes method are described as follows:

- The dataset contains the genetic coding information of 9445 SNP sites of 1000 samples, and the corresponding category label. The base sequence is converted into numerical code information according to the based feature shown in Table 1.
- According to the coding information of 9445 SNP sites, 448 features are selected by chi-square test.
- The naïve Bayes model on 1000 sample are established.
- According to five-fold cross validation, the calculation results of SE, SP, ACC and MCC are obtained, the performance of naïve Bayes prediction model is evaluated.

RESULTS AND DISCUSSION

Due to the presence of redundant information and noise between site sequence data, it is extremely difficult to dig deeper and accurate SNP pathogenic site sequence characteristics. In order to improve the performance of machine learning method, dealing with SNP pathogenic site sequence is an indispensable work. In this paper, according to the principle of chi-square test to reduce the dimension data, remove a lot of redundant information and noise, screening significant SNP pathogenic sites.

The 1000 samples are divided into 500 healthy people and 500 patients with 9445 SNP sites in two datasets. Calculate the frequency of the two datasets in the coding mode of “0”, “1” and “2” at each SNP sites, establish the chi-square test six lattice table, and then perform the chi-square test on the frequency data to obtain the p-value of 9445 SNP sites, the smaller the p-value, the more obvious the difference is between the healthy group and the affected group. In order to better compare the differences of each site, the relative level of each site was observed intuitively. On the basis of maintaining the relative relationship of the data, the comprehensive score of each point p-value was expanded to a certain extent, let p-value take negative logarithm $-Log_{10}(P)$, which are shown in Fig. 3. The possible SNP pathogenic sites are rs2273298, rs2250358, rs7543405, rs932372, rs12036216 and rs9426306.

At the significance level of 0.01, 73 significant SNP pathogenic sites are screened out. When the significance level is 0.05, 447 more significant SNP pathogenic sites are screened. In order to preserve important useful information as much as possible, sexual level is 0.05, in other words, 9445 SNP sites by chi-square test down to 447 SNP pathogenic sites. In process of doing the frequency statistics, finding an abnormal SNP site rs12742921, abnormal frequency statistics table, as shown in Table 3.

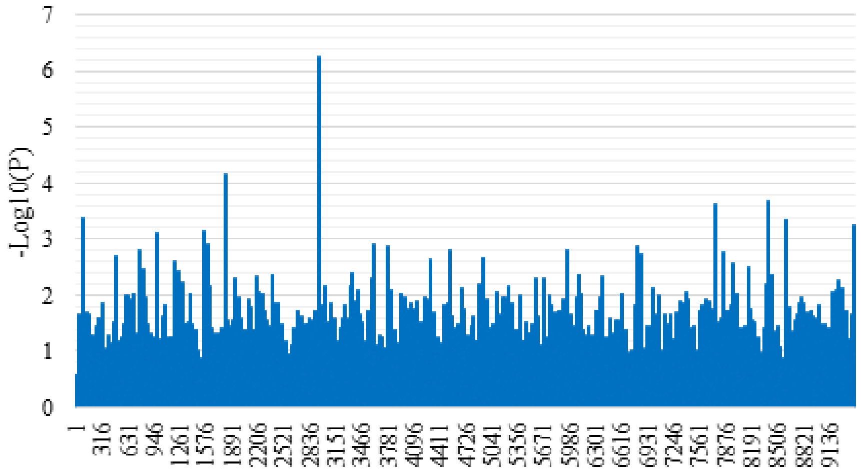

Fig-3: p-value of 9445 SNPs

Table-3: Frequency of SNP rs12742921

	0	1	2
Health	412	88	0
Unhealth	404	91	5

As shown in Table 3, the abnormal SNP site rs12742921 has a significant coding pattern in healthy and unhealthy samples. The frequency of the SNP sites in the healthy sample is “2”, namely, the healthy sample is in the SNP. There are only “0” and “1” encoding, the corresponding non-coding information are “TT” and “TC”, and “CC” does not appear. There are three patterns of coding in unhealthy samples, so the SNP site is likely to be an important factor affecting morbidity and mortality. Although the SNP site of rs12742921 does not pass the chi-square test, the SNP site is predicted to be classified along with the 447 SNP pathogenic sites tested.

In this paper, 1000 sample datasets are used, each of which covers 448 probable SNP pathogenic sites, to establish the naïve Bayes SNP pathogenic site predicted model. The naïve Bayes method is based on the conditional independence hypothesis, in other words, assuming that the influence of an attribute on a given class is independent of other attributes. When the conditional independence assumption is established, the naïve Bayes classification method has the smallest false classification rate, to use in bioinformatics and data mining [31,32]. It generalizes the classifier through the classified training set and classifies the testset using the classifier. The classification of the naïve Bayes method is mainly divided into two stages. The first stage is the learning of the naïve Bayes method that is, constructing the classifier from the training dataset. The second stage is the reasoning of the naïve Bayes classifier, that is, the conditional probability of 448 class nodes is calculated, and the classification data are classified. The naïve Bayes classifier that minimizes the probability of classification errors in a variety of classifiers or minimizes the average risk at a given cost.

In this paper, we use the five-fold cross validation to test the model and get the SE, SP, ACC and MCC are used to evaluate the performance of the model. The SE of the naïve Bayes model is 84.96%, the SP is 84.45%, the ACC is 84.64% and the MCC is 0.6937. The essence is that the ACC of the testset is 84.64%, and the ACC of the healthy sample is 84.96%, the ACC of unhealthy samples is 84.45%, and the MCC of the model is 0.6937.

In order to facilitate the comparison of the predictive ability of the model, we establish the prediction models of SNP pathogenic sites, which is based on AdaBoost, SVM, KNN and RF algorithms. Moreover, these models are compared with the model established in this paper.

AdaBoost is an iterative combination of algorithms [33,34]. In this case, the classifier is used as the basic classifier. The classifier used may be weak at the beginning, namely, the error rate is higher. However, with the running of iteration, the classifier is used to improve the classifier according to the new sample; each iteration is corrected for the misclassification of some of the observations by the previous classifier. It is common way that increases the weight of the observed values in the returned sample. There will be more previous observations of the error, and then form a new classifier into the next iteration, and each iteration of the round of the classifier to give the error rate, the final result is got by the various stages of classification device in accordance with the error rate weighted vote, this is called “adaptive”.

Based on the AdaBoost algorithm, the dataset with the five-fold cross validation has the SE is 73.88%, the specificity is 71.81%, the ACC is 72.78%, and the MCC is 0.4574. The proportion of the samples which have been correctly classified is 72.78%. The proportion of healthy samples which have been correctly classified is 73.88%. The proportion of diseased samples which have been correctly classified is 71.81%. The MCC of the model is 0.4574. Compared with the naïve Bayes method, the AdaBoost algorithm is lower than it, which is 11.86% lower in ACC and 23.63% in MCC.

The naïve Bayes method can accurately predict the data sets, and the SVM is also a widely used machine learning method in data mining technology [35], which can overcome the high dimension of the analyzed data and difficult to reduce the dimension and other issues. In this paper, the SVM is used as a classifier. The sample and its class labels are input into the SVM classifier, to use the five-fold cross validation to test the model to analyze the accuracy of its prediction.

The kernel function has some influence on the prediction results. Therefore, the data are excavated by three kernel functions. The optimal parameters $C = 1$ and $g = 0.001$ are used to determine the ACC of the model and MCC as the verification index and select the highest ACC of the kernel function as a SVM representative. The ACC of the SNP sites is 89.36%, the MCC is 0.5889, the ACC of the polynomial kernel function is 65.54%, the MCC is 0.4225, the ACC of the sigmoid kernel function is 77.08%, and MCC is 0.5431. By comparing with the main results of these three kernel functions, we can see that the radial basis function under the five-fold cross validation method is the highest and the MCC is the highest SE is 81.17%. Compared with the polynomial and sigmoid kernel function, the SP is 77.68%, and its essence is 81.17% for the prediction of healthy samples, and the ACC of the unhealthy samples is 77.68%. The polynomial kernel function and the radial basis function differ by 13.82% in ACC and 0.1664 in MCC. The Sigmoid kernel function differs from the radial basis function by 2.28% in ACC and 0.0458 in the MCC, which indicates that the radial basis function is better in SNP site prediction. Compared with the naïve Bayes method, the SVM is lower than those obtained by the naïve Bayes method, the difference is 5.28% on ACC and 0.1048 on MCC.

KNN are the easiest to implement method in machine learning methods, and are also simpler distance discrimination methods, which have been widely used in the field of bioinformatics. In this paper, k is set to 3, 5, 7, and 9 for five-fold cross validation. When k is 3, ACC is 62.65%, MCC is 0.2607, k is 5, ACC is 63.26%, MCC is 0.2764, k is 7, the ACC is 64.96%, the MCC is 0.3133, and when k is 9, the ACC is 65.56%, and the MCC is 0.3238.

When k is 9, the highest ACC, namely in the five-fold cross validation, 65.56% of the samples are correctly classified. The SE and the SP were also the highest at 77.2% and 54.31%, respectively, indicating that the proportion of the healthy samples was correctly classified is greater than the proportion of the diseased samples being correctly classified, and regardless of the k value, which is better than specificity, the model predicts the phenotype is healthy and the prediction of the phenotype is poor. Compared with the naïve Bayes method, the indexes obtained by the KNN method are lower than those obtained by the naïve Bayes method. It is 19.08% lower in ACC and 0.3699 lower in MCC.

In a variety of machine learning methods, the RF method has good classification performance and high classification accuracy, suitable for a variety of data sets of operations and not sensitive to multiple collinearity so it can be a good prediction of up to thousands explain the problem of variables [36]. The dataset is studied in the random forest classifier, and the model is tested by the five-fold cross validation to obtain the ability of model classification prediction. When the number of variables in the subset of variables randomly selected in the process of growing each tree is zero, the number of variables mtry is 30, ntree is 500, the ACC is 77.78% and the MCC is 0.5689, the ACC for the testset is 77.78%. Compared with the naïve Bayes method, the indexes derived from the RF were lower than those obtained by the naïve Bayes method, and the difference is 6.86% on the ACC and 0.1248 on the MCC.

For ease of comparison, the prediction results of NB, AdaBoost, SVM, KNN and RF are shown in Table 4.

Table-4: Comparison of the prediction results of different methods.

Prediction model	SE (%)	SP (%)	ACC (%)	MCC
NB	84.96	84.45	84.64	0.6937
AdaBoost	73.88	71.81	72.78	0.4574
SVM	81.17	77.68	79.36	0.5889
KNN	77.21	54.31	65.56	0.3238
RF	77.65	78.76	77.78	0.5689

It can be seen from Table 4 that the SNP pathogenic site of the NB model has the highest ACC of 84.53%, the ACC of other models differs from 5% to 20%, 11.86% higher than AdaBoost model, 5.28% higher than the SVM model, 19.08% higher than the KNN model, 6.86% higher than the RF model. MCC is the highest, which is 0.6937; the MCC of other models differs from 0.10 to 0.37, 0.2363 higher than AdaBoost model, 0.1048 higher than SVM model, 0.3699 higher than KNN model, 0.1248 higher than RF model. SE was the highest, which is 84.96%, 11.08% higher than AdaBoost model, the SP was 84.45%, higher than the KNN model by 30.14%. The results show that the SNP pathogenic site activity of NB model is the best. In order to better compare the different machine learning methods, the relative of each method was observed intuitively which are shown in Fig. 4.

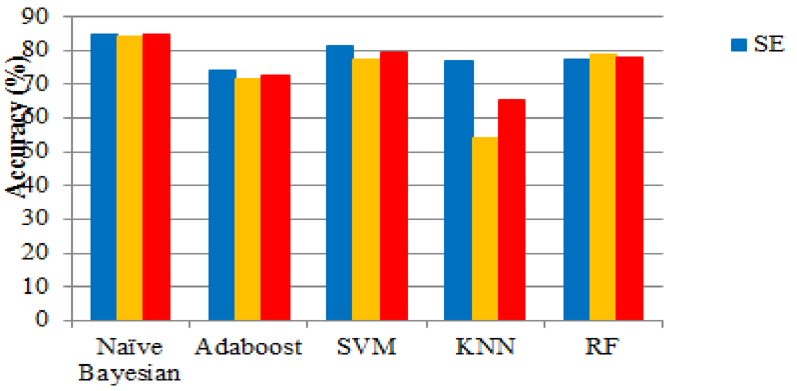

Fig-4: The prediction results of five machine learning methods.

As shown in Fig. 4, the NB model has the highest ACC, and its SE, SP and MCC are the highest, followed by RF and SVM model. When the SVM method deals with large-scale data, the time complexity and spatial complexity increase linearly with the increase of the data volume, and the parallel training is difficult, so the model performance is worse than the naïve Bayes.

The prediction effect on the AdaBoost model and the KNN model is not ideal, possibly because the 448 SNP sites extracted by the chi-square test are not sufficient to fully interpret all the site data. The AdaBoost algorithm only needs to set the number of iterations, in order to get the exact classification model, the setting of a large number of iterations does not stop running, resulting in the iterative late sub-classifier to improve the generalization performance of the classifier is very small, and then there is a risk of fitting.

There is a large error in the classification prediction using KNN, it makes the prediction result of the KNN model poor. Compared with the predictive performance of other models, naïve Bayes has the smallest error classification rate and the logic is simple. The model needs to estimate very few parameters and easy to implement. Besides the calculation is fast, robust, stable performance, for different characteristics of the data classification performance is not very different, that is, the robustness of the model is better. The experimental results show that the constructed naïve Bayes model can favorably predict the classification.

CONCLUSION

With the rapid development of information technology, the era of big data has come. There is an urgent need to develop a more convenient and effective tool to classify the collection of massive bioinformatics data quickly and accurately, in order to obtain the information which we need [37,38]. This paper conducts study of SNP site data and presents a new method for SNP pathogenicity prediction based on naïve Bayes method. It can get the ACC and MCC by using five-fold cross validation. Classification performance of five machine learning methods is compared and shows that the model can be used to predict the disease effectively according to the locus of genetic disease, which is an ideal SNP method for predicting disease. At present, the researchers have used a variety of machine learning methods for SNP pathogenic site prediction and identification, including the KNN, neural network, DT and SVM and other supervised learning methods for genetic disease SNP pathogenic site Classification prediction is still in initiation and exploration stage, and the use of machine learning methods to predict SNP pathogenic site has become a hotspot in the field of bioinformatics. Deep learning is a new machine learning method, and the use of it to predict SNP pathogenic site will be the next step in the future.

REFERENCES

1. El-Aksher SH, Sherif HS, Khalil MH, El-Garhy HA, Ramadan S. Molecular analysis of a new synthetic rabbit line and their parental populations using microsatellite and SNP markers. Gene Reports. 2017 Sep 1;8:17-23.
2. Kuang M, Wei SJ, Wang YQ, ZHou DY, Ma L, Fang D, Yang WH, Ma ZY. Development of a core set of SNP markers for the identification of upland cotton cultivars in China. Journal of Integrative Agriculture. 2016; 15(5):954-962.
3. Liao B, Li X, Zhu W, Li RF, Wang SL. Multiple ant colony algorithm method for selecting tag SNPs. Journal of Biomedical Informatics. 2012; 45(5):931-937.
4. Chuang LY, Yang CS, Ho CH, Yang CH. Tag SNP selection using particle swarm optimization. Biotechnology Progress. 2010; 26(2):580-588.
5. Yu B, Li S, Liu H. A hybrid gene selection method for tumor classification based on genetic algorithm and support vector machine. Journal of Computational & Theoretical Nanoscience. 2015; 12(11):4730-4735.
6. Sun YJ, Todorovic S, Li J. Unifying multi-class AdaBoost algorithms with binary base learners under the margin framework. Pattern Recognition Letters. 2007; 28(5):631-643.
7. Yu B, Zhang Y, Zhao L. Cancer classification by a hybrid method using microarray gene expression data. Journal of Computational & Theoretical Nanoscience. 2015;12(10):3194-3200.
8. Sadeghi BHM. A BP-neural network predictor model for plastic injection molding process. Journal of Materials Processing Technology. 2000; 103(3):411-416.
9. İLhan İ, Tezel G. A genetic algorithm–support vector machine method with parameter optimization for selecting the tag SNPs. Journal of biomedical informatics. 2013 Apr 1;46(2):328-40.
10. Halperin E, Kimmel G, Shamir R. Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics. 2005; 21(Suppl 1):195-203.
11. Yu B, Lou L, Li S, Zhang Y, Qiu W, Wu X, Wang M, Tian B. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. Journal of Molecular Graphics and Modelling. 2017 Sep 1;76:260-73.
12. Lee PH, Shatkay H. BNTagger: improved tagging SNP selection using Bayesian networks. Bioinformatics. 2006; 22(14):211-219.
13. Patil N. Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21. Science. 2001; 294(5547):1719-1723.
14. Mahdevar G, Zahiri J, Sadeghi M, Nowzari-Dalini A, Ahrabian H. Tag SNP selection via a genetic algorithm. Journal of Biomedical Informatics. 2010; 43(5):800-804.
15. Phuong TM, Lin Z, Altman RB. Choosing SNPs Using Feature Selection. Journal of Bioinformatics and Computational Biology. 2006; 4(2):241-257.
16. Halldorsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S. Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Research. 2004; 14(8):1633.
17. Pele O, Werman M. The quadratic-chi histogram distance family. InEuropean conference on computer vision. Springer, Berlin, Heidelberg. 2010;(5):749-762.
18. Lee CH. A gradient approach for value weighted classification learning in naive Bayes. Knowledge-Based Systems. 2015 Sep 1;85:71-9.
19. Jiang LX, Cai ZH, Zhang H, Wang DH. Naïve Bayes text classifiers: a locally weighted learning approach, Experimental & Theoretical Artificial Intelligence. 2013; 25(2):273-286.

20. Cao Y, Miao Q , Liu JC, Gao L. Advance and Prospects of AdaBoost Algorithm. Acta Automatica Sinica. 2013; 39(6):745-758.
21. Vapnik V. The nature of statistical learning theory. Springer. 1995.
22. Chang CC, Lin CJ. A Library for Support Vector Machines. 2011; 2(3):1-27.
23. Yu B, Li S, Chen C, Xu J, Qiu W, Wu X, Chen R. Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition. Chemometrics and Intelligent Laboratory Systems. 2017 Aug 15;167:102-12.
24. Yu B, Li S, Qiu WY, Chen C, Chen RX, Wang L, Wang MH, Zhang, Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. Oncotarget. 2017; 8(64):107640-107665.
25. Mikhchi A, Honarvar M, Kashan NEJ, Zerehdaran S, Aminafshar M. Analyses and comparison of K-nearest neighbour and AdaBoost algorithms for genotype imputation. Research Opinions in Animal & Veterinary Sciences. 2015; 5(7):295-299.
26. Breiman L, Random forests. Machine Learning. 2001; 45(1):5-32.
27. Vincent B, Gilles L, Pierre G, Wehenkel L. Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. PLoS ONE. 2014; 9(4):e93379.
28. Meng, YA, Yu Y, Cupples L, Farrer LA, Lunetta, KL. Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics. 2009; 10(1):78.
29. Goldstein A, CPolley E, Briggs F. Random forests for genetic association studies, Statistical Applications in Genetics and Molecular Biology. 2011;10(1):1-34.
30. Qiu W, Li S, Cui X, Yu Z, Wang M, Du J, Peng Y, Yu B. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. Journal of theoretical biology. 2018 Aug 7;450:86-103.
31. Griffis IC, Allendorfer JB, Szaflarski JP. Voxel-based gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. Journal of Neuroscience Methods. 2016; 257:97-108.
32. Zhang H, Kang YL, Zhu YY, Liang JY. Novel naïve Bayes classification models for predicting the chemical ames mutagenicity. Toxicology in Vitro. 2017; 41:56-63.
33. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning. 2000; 40(2):139-157.
34. Mukherjee I, Schapire E. A theory of multiclass boosting. Journal of Machine Learning Research. 2013; 14(1):437-497.
35. Claesen M, Smet FDE, Suykens JAK. EnsembleSVM: a library for ensemble learning using support vector machine. Journal of Machine Learning Research. 2014; 15(1):141-145.
36. Goldstein BA, Polley EC, Briggs F. Random forests for genetic association studies. Statistical Applications in Genetics and Molecular Biology. 2011; 10(1):1-34.
37. Yu B, Zhang Y. The analysis of colon cancer gene expression profiles and the extraction of informative genes. Journal of Computational and Theoretical Nanoscience. 2013; 10(5):1097-1103.
38. Wong TT. Generalized dirichlet priors for naïve Bayesian classifiers with multinomial models in document classification. Data Mining and Knowledge Discovery. 2014; 28(1):123-144.