Identification of Gut Flora Based on Robust Support Vector Machine

Xinxin Wang¹*, Jingjing Liu¹, Liya Ma¹
¹Heze Homemaking Professional College, Heze, Shandong, China
*email: bosi@nuist.edu.cn

Abstract. Gut flora parasitize the human gastrointestinal tract to maintain normal physiological functions, but can also lead to a variety of diseases and affect mental health. The relationship between gut flora and human health has attracted increasing attention and become a popular research hotspot at present. However, biological data are often characterized by large size and high dimensionality, leading to limited ability of traditional statistical-based methods to handle these data and make them difficult to analyze. In this paper, we use a support vector machine (SVM) model to perform data mining analysis on the Gut flora dataset. Specifically, we build a robust SVM model to compare the gut flora of obese and healthy people based on the public database of the American Gut Project, analyze the characteristics of the gut flora of obese people, and set up a machine learning model to predict the obesity status of people based on the gut flora to provide a theoretical basis for obesity intervention based on the gut flora. Different from the traditional support vector machine model, we use ramp loss to calculate the sample loss, avoid the influence of noise data, and get more accurate experimental results. The experimental results reveal the characteristics of the gut flora of obese people, apply machine learning to obesity prediction, and provide new research ideas and theoretical basis for precision diet and precision medicine.

1. Introduction
Humans and microbial communities have formed an ingenious symbiotic relationship over a long period of evolution, and these microorganisms are present in large numbers in the human skin, mouth, and intestines, affecting the body’s metabolism. Because of this, the study of the composition of microbial communities has become an important aspect of the medical field.

The gut flora is a collective term for the collection of microorganisms that live in symbiosis with humans in the human gastrointestinal tract. The gut flora has a significant role in human health and has attracted more and more attention from researchers in recent years. Generally speaking, the number of microorganisms in the normal human gut is about $10^{11} \sim 10^{12}$. These microorganisms are acquired from the surrounding environment at birth and then colonize the human intestine, where they are closely related to the normal functioning of the body. The number of microorganisms in the intestine is about 10 times greater than the total number of cells in the human body. The gut flora is divided into three major groups: probiotics, harmful bacteria and neutral bacteria. Their ratio maintains a balance in a healthy human body, and because of this, the normal metabolism and physiological functions of the human body can be maintained. In recent years, many studies have shown that there is a link between the gut flora and the digestive, nutritional, metabolic and immune aspects of the host, and that there is a link between disorders of the intestinal flora and many diseases. For example, a large number of...
studies have shown a correlation between intestinal flora disorders and irritable bowel syndrome [1], inflammatory bowel disease [2], colon cancer [3], obesity [4] and diabetes [5].

In this paper, we study the potential link between obesity and gut flora. Many studies have shown that disturbances in the intestinal flora may be an important cause of obesity [6]. Some studies have found a lower Firmicutes/Bacteroidetes ratio in the gut of obese people [7], yet some studies have found the opposite, such as the study by Mai et al. [8] did not find an association between Firmicutes/Bacteroidetes ratio and BMI. Therefore, further studies are needed to investigate the changes in the abundance of Firmicutes and Bacteroidetes in the intestinal flora of obese people. At the genus level, Schwiertz et al [9] reported a decrease in the relative abundance of Methanobrevibacter in the intestine of obese people compared to healthy people.

With the gradual arrival of the era of medical informatization, the huge amount of medical data accumulated by medical institutions finally has a place to be used. While traditional research methods represented by statistics are difficult to find useful information from the huge amount of medical data, data mining technology can dig out valuable information in the huge amount of data, which plays an important role in the research of disease causes and so on. In this paper, we use a SVM model to perform data mining analysis on the Gut flora dataset. Specifically, we build a robust SVM model to compare the gut flora of obese and healthy people based on the public database of the American Gut Project, analyze the characteristics of the gut flora of obese people, and set up a machine learning model to predict the obesity status of people based on the gut flora to provide a theoretical basis for obesity intervention based on the gut flora. Different from the traditional support vector machine model, we use ramp loss to calculate the sample loss, avoid the influence of noise data, and get more accurate experimental results.

2. Proposed Approach

2.1. Standard Support Vector Machine

In supervised learning, Support Vector Machine (SVM) is a powerful classification method that is widely used to separate data by maximizing the margin between two classes. Given a training set \( D = \{(x_i, y_i)\}_{i=1}^{n} \) with \( n \) samples, where \( x_i \in \mathbb{R}^d \) and \( y_i \in \{+1, -1\} \), the standard SVM learns a decision hyperplane \( f(x_i) = w^T x_i + b \).

In order to maximize the distance between the samples closest to the decision boundary and the decision boundary, the standard SVM solves the following problem:

\[
\min_{w, b} \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) > 1 \quad (1)
\]

By introducing Lagrange multiplier \( \alpha \), we solve the dual problem of Eq. (1) as follows:

\[
\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad \text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0 \quad (2)
\]

We can solve Eq. (2) by quadratic programming. Suppose that the optimal solution obtained by quadratic programming is \( \alpha^* = [\alpha_1^*, \alpha_2^*, \cdots, \alpha_n^*] \), the optimal solution is as follows:

\[
w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i, b^* = -\frac{1}{2} w^* (x_i + x_j) \quad (3)
\]

where \( x_i \) and \( x_j \) are support vector. However, in reality, many datasets have a few samples that are linearly indistinguishable, a situation that can lead to the algorithm not finding the optimal
classification hyperplane. In this case, we can transform the optimization objective by introducing slack variables $\xi$ into:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(w^T x + b) > 1 - \xi_i, \xi_i \geq 0$$

(4)

where $C$ is penalty factor, which is used to control the degree of misclassification. Similarly, we use the dual problem to solve Eq. (4). By introducing Lagrange multiplier $\alpha$, we solve the dual problem of Eq. (4) as follows:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \quad \text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C$$

(5)

We call this new type of SVM a soft margin SVM. The soft margin criterion in SVM allows the samples to be misclassified at a certain loss to deal with the non-linearly separable data. In the case of the intestinal flora data structure in this paper, it is not common to have linearly divisible problems in the process of practical application. If we encounter a nonlinear problem, we need to make further changes to the linearly divisible SVM. Specifically, we take the linearly indistinguishable samples in the original space and project them into the high-dimensional space according to certain rules, so that we can find the decision hyperplane in the high-dimensional space. There may still be some cases where the sample points are linearly indistinguishable after the projection, and we can still use the soft margin SVM to handle these cases. In solving the dual problem in the original space is required to solve a quadratic programming problem using the dual, and the solution requires computing the dot product of the sample point vectors, which is similar in the high-dimensional space. We first transform to the dual problem, but solving the dot product is very difficult in the high-dimensional space. Therefore, we solve this problem in the form of a kernel function. The kernel function needs to satisfy the Mercer condition:

$$K(x_i, x_j) = \phi(x_i) \phi(x_j)$$

(6)

When transforming to a high-dimensional space to find classification surfaces, the kernel function solves the complex dot product calculation, and the solution becomes more convenient:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad \text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C$$

(7)

2.2. Robust Support Vector Machine

Hinge loss is a common loss function in soft margin SVM

$$H_s(z) = \max(0, s - z)$$

(8)

where $s$ indicates the position of hinge point and the elbow at $s = 1$ indicates the point at which $y_i f(x_i) = y_i (w \cdot \phi(x_i) + b) = 1$. In general, the standard convex SVM limits the influence of any single training sample, since $0 \leq \alpha \leq C$ by Eq. (5), as Figure 1 blue dotted line illustrates.

This results in the standard convex SVM are still inappropriately drawn toward outliers samples, and all misclassified training samples become support vectors. Loss clipping is a convenient method to obtain a bounded loss from a convex loss. We clip the hinge loss to get the ramp loss:

$$R_s(z) = \min(1 - s, H_s(z)) = H_s(z) - H_s(z)$$

(9)

where $s \leq 0$. The ramp loss can be decomposed into a convex hinge loss and a concave loss. After using the ramp loss, the original formulation Eq. (4) can be transformed as follows:
\[
\min_{w,b} \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^{n} H_{i}(y_{i}f(x_{i})) - C \sum_{i=1}^{n} H_{i}(y_{i}f(x_{i}))
\]

(10)

where \(a, v\) indicates real-valued convex functions. As shown in [9], the objective function is expressed as the form of a difference of convex (DC) functions. However, the non-convexity of the minimization objective (13) can lead to significant difficulties in optimization. We propose instead to optimize Eq.(13) using the Concave-Convex Procedure (CCCP).

\[
[w, b]v(w, b) = -\sum_{i=1}^{n} \mu_{i}y_{i}f(x_{i}), \text{ where } \mu_{i} = \begin{cases} C & \text{if } y_{i}f(x_{i}) < s, \\ 0 & \text{otherwise.} \end{cases}
\]

Fig. 1 Margin loss function

2.3. CCCP for robust support vector machine

In this paper, we employ the CCCP to solve the Eq.(13). The CCCP procedure decomposes a non-convex cost function into a combination of convex parts and performs optimization on the difference of these convex functions. Especially, the idea of CCCP is to linearize the concave part of (13) around a solution obtained in the current iteration so that \(o(w, b) - [w, b]v(w, b)\) is convex in \((w, b)\). We compute the \([w, b]v(w, b)\) as follows:

\[
[w, b]v(w, b) = -\sum_{i=1}^{n} \mu_{i}y_{i}f(x_{i}), \text{ where } \mu_{i} = \begin{cases} C & \text{if } y_{i}f(x_{i}) < s, \\ 0 & \text{otherwise.} \end{cases}
\]

We can obtain the primal convex inner loop problem for (13) based on the CCCP algorithm, and the corresponding dual convex inner loop problem is defined as follows:

\[
\min_{a} \frac{1}{2} \alpha^{T}H\alpha - y^{T}\alpha \quad \text{s.t.} \quad \sum_{i=1}^{n} \alpha_{i} = 0; A_{i} \leq \alpha_{i} \leq B_{i}, \quad \text{where} \quad \begin{cases} A_{i} = \min(0, C_{i}) - \mu_{i}y_{i} \\ B_{i} = \max(0, C_{i}) - \mu_{i}y_{i} \end{cases}
\]

(12)

where \(H\) is a positive semi-definite matrix with \(H_{ij} = K(x_{i}, x_{j}) = \langle \phi(x_{i}), \phi(x_{j}) \rangle\) for all \(1 \leq i, j \leq n\), \(K(x_{i}, x_{j})\) is the kernel function.

3. Experiments and analysis

3.1. Experimental setup

To reveal the association between gut flora and human health obesity status, we used support vector machine, robust support vector machine, random forest (RF), gradient boosting decision tree (GBDT) and BP neural network to build classification models, respectively. RF is a collection of many decision trees, so the important parameters affecting random forest are the number of decision trees and the maximum number of features that can be used by a single decision tree, The parameters to be adjusted for GBDT are the number of decision trees and learning rate. Different from random forest, the performance of gradient lifting regression tree does not strongly depend on the number of decision trees. BP neural network (BP) selects three-layer structure, and its main parameter is the number of nodes in the hidden layer. We used grid search algorithm to find the appropriate hyperparameters for
each comparison algorithm. We traverse all parameter combinations in the hyperparameter space of the five comparison algorithms and use five-fold cross-validation (CV) to ensure the reliability of the results and compare the test set accuracy and AUC values before and after the models adjust their parameters. We used grid search algorithm to find appropriate hyperparameters for each comparison algorithm. We traverse all parameter combinations in the hyperparameter space of the five comparison algorithms and use five-fold CV to ensure the reliability of the results and compare the test set accuracy and AUC values before and after the models adjust their parameters. The data used were obtained from the public dataset of the American Gut Project, from which we screened the final valid sequenced sequences of the gut flora artifacts above 1250. From these, we then selected 1655 healthy and 898 obese samples based on BMI. The healthy population was defined as having a BMI between 18.5 and 25.0 and no history of antibiotic medication, inflammatory bowel disease or diabetes within one year; the obese population was defined as having a BMI above 30. alpha diversity index, beta diversity index and OTU tables were derived from the QIIME-based analysis platform of the American Gut Project.

3.2. Results and discussion
Table 1 gives the result of Accuracy and AUC after grid searching. Fig. 2 shows the accuracy before and after grid searching. Fig. 3 shows the AUC before and after grid searching. Based on the analysis results, we used a machine learning approach to construct an obesity prediction model. Through systematic network parameter search, we found that RSVM based on intestinal flora had good performance. Generally speaking, the performance of the model is better when the AUC value is greater than 0.75, while the AUC value of the model with RSVM in this study reached 0.778, and the test accuracy reached 0.672. Compared with SVM, the result of RSVM is improved by more than 5%. The model performance of RSVM is better than the other four models in this study, which reflects the effectiveness of our algorithm. Compared with traditional SVM, RSVM can effectively suppress the influence of noise in the data and greatly improve the accuracy of classification results. These results indicate that the gut flora can predict human health obesity status, revealing another function of the gut flora. Not only that, this model also provides an important reference for obesity intervention based on gut flora. The above series of studies not only demonstrate the great potential of gut flora in disease prediction, but also provide a solid theoretical foundation for future precision diet and precision medicine.

| Method | BP  | GBDT | RF  | SVM | RSVM |
|--------|-----|------|-----|-----|------|
| Accuracy | 0.632 | 0.654 | 0.643 | 0.629 | 0.672 |
| AUC | 0.631 | 0.754 | 0.743 | 0.729 | 0.778 |

Fig. 2 The accuracy before and after grid searching
4. Conclusion

We use a SVM model to perform data mining analysis on the Gut flora dataset. Specifically, we build a robust SVM model to compare the gut flora of obese and healthy people based on the public database of the American Gut Project. Different from the traditional support vector machine model, we use ramp loss to calculate the sample loss, avoid the influence of noise data, and get more accurate experimental results. The experimental results reveal the characteristics of the gut flora of obese people, apply machine learning to obesity prediction, and provide new research ideas and theoretical basis for precision diet and precision medicine.

References

[1] Ghoshal U C, Shukla R, Ghoshal U, Gwee K A, Ng S C and Quigley E M 2012 The gut microbiota and irritable bowel syndrome: friend or foe? International journal of inflammation

[2] Manichanh C, Borruel N, Casellas F and Guarner F 2012 The gut microbiota in IBD Nature reviews Gastroenterology & hepatology 9(10) pp 599-608

[3] Arthur J C, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis J M, Fan T J and Jobin C 2012 Intestinal inflammation targets cancer-inducing activity of the microbiota science 338(6103) pp 120-123

[4] Clarke S F, Murphy E F, Nilaweera K, Ross P R, Shanahan F, O’Toole P W and Cotter P D 2012 The gut microbiota and its relationship to diet and obesity: new insights Gut Microbes 3(3) pp 186-202

[5] Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F and Wang J 2012 A metagenome-wide association study of gut microbiota in type 2 diabetes Nature 490(7418) pp 55-60

[6] Reinhardt C, Reigstad C S and Bäckhed F 2009 Intestinal microbiota during infancy and its implications for obesity Journal of Pediatric Gastroenterology and Nutrition 48(3) pp 249-256

[7] Ley R E, Turnbaugh P J, Klein S and Gordon J I 2006 Human gut microbes associated with obesity nature 444(7122) pp 1022-1023

[8] Mai V, McCrary Q M, Sinha R and Glei M 2009 Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers Nutrition Journal 8(1) pp 1-10

[9] Sriperumbudur B K and Lanckriet G R 2019 On the Convergence of the Concave-Convex Procedure In Nips 9 pp 1759-1767