Application of Machine Learning in Cigarette Quality Evaluation Method

Jingju Lin¹, Haolin Guo²

¹Longyan Tobacco Industry Co., Ltd. Fujian, Longyan, 364000
²ShaoGuan Cigarette Factory Of China Tobacco GuangDong Industrial Co.Ltd, ShaoGuan, 512000

Abstract: In the evaluation of cigarette quality, it usually depends on the sense organs of evaluation experts. But due to the influence of many subjective and objective factors, the accuracy of evaluation results is often difficult to guarantee. Therefore, this paper proposes a method of cigarette quality evaluation based on SVM and BP neural network. The experimental results show that the method can establish a nonlinear mapping relationship between the measured values of cigarette chemical components and the evaluation values of expert quality, and reflect the preference information and reasoning mechanism of evaluation experts. So the model is an objective and reliable method for evaluating the internal quality of tobacco after knowing the chemical composition test data of tobacco.

1. Introduction
At present, the tobacco industry mainly relies on the sensory evaluation of smoking experts to evaluate the quality and style of cigarettes. Sensory evaluation mainly depends on human physiological and psychological conditions, which is a delicate technology. However, such assessment results are often affected by the knowledge structure, experience, emotion, environmental conditions and other factors of experts, so the reliability of the assessment results is difficult to guarantee. In addition, long-term engagement in the work of evaluation and absorption is harmful to the physical and mental health of evaluation and absorption experts. Therefore, many tobacco workers have been exploring the relationship between the chemical composition of tobacco and its internal quality, trying to directly use the chemical composition of tobacco to determine the internal quality of cigarette products. Most of these studies use the traditional mathematical statistical methods for analysis. However, the chemical components of tobacco are extremely complex, and the relationship between their sensory stimulation and human subjective feelings is extremely complex and subtle, so it is impossible to establish a mathematical model to determine the relationship between the chemical components of tobacco and the internal quality of cigarette products. The mathematical statistics method can only analyze the relevant factors that affect the quality of cigarettes and give the degree of influence, but not the evaluation results directly.

Support vector machine and BP neural network as common machine learning algorithms, can ensure that the learning machine has good generalization ability. The solution of SVM is transformed into a quadratic programming problem, which guarantees the global optimization of the algorithm and solves the local minimum problem that cannot be avoided by neural network. In this paper, based on the historical data of cigarette artificial evaluation, a new way to solve the problem of cigarette quality evaluation is proposed by using SVM to establish the relationship between the chemical composition of tobacco and the internal quality of cigarette products.

2. Machine Learning Algorithm

2.1 Support Vector Machine
Support vector machine (SVM) shows significant advantages in solving small sample, nonlinear and high-dimensional pattern recognition problems. The SVM algorithm can construct good classification rules in a very high-dimensional space, and provide a unified theoretical framework for the classification algorithm.

For the sample set \[ G = \{(x_i, d_i)\}_{i=1}^n \] (\( x_i \) is Input vector; \( d_i \) is Corresponding target value; \( n \) is Number of samples), the SVM regression function is:

\[
y = f(x) = wO(x) + b
\] (1)

In the formula, \( O(x) \) is a high-dimensional feature space obtained by nonlinear transformation from the input space \( x \). The coefficients \( k \) and \( b \) are obtained by minimizing the risk function:

\[
R_{SVM}(C) = C \sum_{i=1}^{n} L(x_i, y_i) + \frac{1}{2} \|w\|^2
\] (2)

\[
L(x, y) = \begin{cases} 
|d - y| - X, & |d - y| \geq X \\
0, & \text{others}
\end{cases}
\] (3)

In the risk function shown in Formula (2), \( C \sum_{i=1}^{n} L(x_i, y_i) \) represents empirical risk, which is measured by the insensitive loss function \( X \) in Formula (3). The loss function allows us to describe the regression function in Formula (1) with sparse data. \( \frac{1}{2} \|w\|^2 \) is a regularization term. \( C \) is the regularization parameter, which controls the complexity and generalization ability of the regression function. \( X \) is visually compared to a pipeline, which is a design parameter directly related to the accuracy of function estimation. Parameters \( C \) and \( X \) need to be set in advance.

When constraints are unrealizable, introduce slack variables \( Y_i \) and \( Y_i^* \), then Formula (2) can be rewritten as:

\[
\begin{align*}
\min R_{SVM}(w, Y) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (Y_i + Y_i^*) \\
\text{s.t.} & \quad d_i - wO(x_i) - b_i \leq X + Y_i \\
& \quad wO(x_i) + b_i - d_i \leq X + Y_i^*, \ Y_i^* \geq 0
\end{align*}
\] (4)

Use the Lagrange multiplier method to solve this quadratic programming problem with linear inequality constraints, and obtain the regression function of Formula (a) as:

\[
f(x, T, T^*) = \sum_{i=1}^{n} (T_i - T_i^*) K(x, x_i) + b
\] (5)

2.2 BP Neural Network

BP neural network is a multilayer feedforward network, mainly by simulating the feedback behaviour of neurons in the human brain to respond to external signal stimuli. BP neural network is mainly composed of three layers of network, input layer, hidden layer and output layer. Each layer contains multiple parallel neuron signals. The BP neural network mainly through the forward propagation of the signal and the reverse adjustment of the error, find the weights of each neuron connected to the different levels, to build the BP neural network model. In the forward propagation process, the signal flows through the input layer and the hidden layer and flows to the output layer. In the error back propagation, the error between the output layer and the expected output is used as the source information, and the connection weights and threshold between each neuron are dynamically adjusted. This
training is repeated many times until the output of the model approximated the desired output. BP neural network algorithm has the characteristics of self-regulation, high self-study, the most important thing is to overcome the incompleteness of information and can make good use of many complex nonlinear relational data to quickly model. Through a specific reverse adjustment algorithm, we can autonomously learn a set of representative connection weights and thresholds and have good robustness and generalization. In this paper, the Sigmoid function is chosen as the transfer function of the hidden layer. Taking the calculation of the hidden layer of the first level as an example, the formula is as follows:

$$a_{0}^{(2)} = g(\theta_{10}^{1}x_{0} + \theta_{11}^{1}x_{1} + \theta_{12}^{1}x_{2} + \cdots + \theta_{1n}^{1}x_{n})$$

$$a_{1}^{(2)} = g(\theta_{20}^{1}x_{0} + \theta_{21}^{1}x_{1} + \theta_{22}^{1}x_{2} + \cdots + \theta_{2n}^{1}x_{n})$$

$$\cdots$$

$$a_{m}^{(2)} = g(\theta_{m0}^{1}x_{0} + \theta_{m1}^{1}x_{1} + \theta_{m2}^{1}x_{2} + \cdots + \theta_{mn}^{1}x_{n})$$

In this, a represents the exciting value of each neural node. The function g represents the Sigmoid function:

$$g(x) = \frac{1}{1 + e^{-x}}$$

Finally, the output value is calculated. The output layer is determined by the neurons in the hidden layer connected to it.

3. Cigarette Quality Evaluation Model Based on Machine Learning Algorithm

3.1 Data Preprocessing and Classification

Accurate decision-making depends on high-quality data. Cigarette quality assessment first requires data preprocessing. This paper focuses on data integration, data cleaning, data transformation and data reduction of cigarette quality data.

a) Data cleaning: data cleaning refers to eliminating noise and inconsistent data by filling in missing values, smoothing noise and identifying or deleting outliers.

b) Data integration: data integration refers to the integration of data from multiple data sources into a consistent storage. When matching the attributes of one database to another during integration, we should pay special attention to the structure of the data, which ensures that attribute function dependencies and reference constraints in the source system match those in the target system. Careful integration of data from multiple data sources can help reduce and avoid redundancy and inconsistency in the result data set, which helps to improve the accuracy and speed of the subsequent mining process.

c) Data conversion: it is to transform data into a form suitable for mining, mainly including data smoothing, aggregation, generalization, normalization and attribute construction.

d) Data reduction: Data reduction technology can be used to get the reduction representation of data sets, which is much smaller, but still close to maintaining the integrity of the original data. Data reduction strategies include data cube aggregation, attribute subset selection, dimension reduction, discretization and concept hierarchical generation.

We use the support vector machine theory to classify the quality data of cigarette. SVM algorithm was originally designed for binary classification problem. When dealing with multi class problems, it is necessary to construct a suitable multi class classifier. At present, there are two main methods to construct multi class classifiers: one is direct classification, that is, to directly modify the objective function, combine the parameter solution of multiple classification surfaces into an optimization problem, and realize multi classification by solving the optimization problem; the other is indirect classification, that is, to realize the construction of multiple classifiers by combining multiple two classifiers.

3.2 Model Indicator Setting

(1) Sensory quality evaluation index of cigarette
In this paper, gloss, aroma, coordination, impurity, irritancy and aftertaste are used to measure the sensory quality of cigarettes. For these six evaluation indexes, we define the grades according to the rating range and qualitative description of different grades is given as Table 1.

Table 1. Grade and description of cigarette sensory quality

| Grade  |
|--------|
| Gloss | Aroma | Coordination | Impurity | Irritancy | Aftertaste |
| 8~10  | Oily | Rich and full | - | - | - |
| 5~8   | Relatively oily | Substantial but rough | - | - | - |
| 1~4   | Relatively dim | Light | Inadequate coordination | - | - | - |
| -10~8 | - | - | - | Mixed gas | Pungent | General |
| -7~5  | - | - | - | Slightly mixed gas | Slightly pungent | Relatively comfortable |
| -4~1  | - | - | - | Bo mixed gas | Non-pungent | Pure and comfortable |

For small sample sets, we use the gray analytic method to assign values to each indicator, and finally determines the impact factor based on the results obtained.

(2) Quality influencing factor index

There are many factors that affect the sensory quality of cigarettes, we selected sugar content, nitrogen content, nicotine content, protein content and Shmuck value as the influencing factor index, as shown in Table 2.

Table 2. Quality influencing factor index

| Factors        | Influence on cigarette quality |
|----------------|--------------------------------|
| Sugar content  | Appropriate sugar content can soften smoke, reduce impurity gas and irritation; too high sugar content will make the smoke flat and reduce the aroma. |
| Nitrogen content | Appropriate nitrogen content will contribute a certain aroma, but it will also cause alkaline substances in the combustion process, producing spicy and bitter taste. |
| Nicotine content | Affects the end of the cigarette. |
| Protein content | Protein content affects cigarette aroma; excess protein will produce a bad smell like burning feathers. |
| Shmuck value   | Affects the PH balance of cigarettes; under certain conditions, the higher the value, the better. |

3.3 Model Building

Under the condition of different physical and chemical indexes, different sensory quality grades may appear in cigarettes. However, under the same condition of conventional physical and chemical indexes, different quality grades may be produced. Therefore, it is necessary to adjust the conventional physical and chemical indexes of the offset in the formula to achieve or maintain the required quality level.

Assume $Y_{ij} = \theta_{ij} + \omega_{ij}$, $\theta_{ij} \sim N(\mu, \sigma$)

In the formula, $Y_{ij}$ is the deviation value between the measured data and the theoretical data of the sensory quality characteristics of the cigarette; $\theta_{ij}$ is the system and measurement error; $i$ is the number of times the process deviation occurs; $j$ is the $j$-th product or sample group affected by the $i$-th influencing factor.

Assume $u_{ij}$ is the adjustment measures taken for the sensory quality characteristics of the $j$-th product or product sample group affected by the $i$-th influencing factor, then the relative adjustment measures can be obtained as:

$$U_{ij} = u_{ij} - u_{ij-1}$$  \hspace{1cm} (9)

$$\theta_{ij} = \theta_{ij} - U_{ij-1}$$  \hspace{1cm} (10)
Then,

\[ Y_i = \theta_i + \omega_i = \theta_i + u_{ij} + \omega_j \]  \tag{11}

If \( \theta_i \) is known, then assume \( u_{ij} = -\theta_j \) to complete the adjustment, but due to the existence of the evaluation and absorption error, it cannot be directly obtained through measurement, so the key of process adjustment is to estimate unknown parameters.

4. Implementation and Example of the Evaluation Algorithm

4.1 Algorithm Implementation
According to the above SVM learning algorithm and the SVM-based cigarette quality evaluation model, we present the algorithm steps of the SVM method for cigarette quality evaluation as follows:

a) Determine the index matrix of cigarette quality evaluation as \( A \);

b) Transform the index matrix \( A \) into the normalized matrix \( R \);

c) Constitute the learning sample set \( G = \{ (x_i, u_i) \}_{i=1}^{l} \) based on cigarette sample \( x_i = (r_{i1}, r_{i2}, \cdots, r_{im}) \) and expert evaluation value \( u_i \). Randomly select samples from sample set \( G \) to form the training set and verification set of SVM learning.

d) Select kernel function and determine appropriate parameters, including SVM model parameters and kernel function parameters.

For the evaluation of cigarette quality, the criterion of parameter selection is:

\[
\begin{align*}
\text{minMAE} &= \frac{1}{l} \sum_{i=1}^{l} |u_i^e - u_i^{SVM}| \\
\text{MSE} &= \frac{1}{l} \sum_{i=1}^{l} (u_i^e - u_i^{SVM})^2 
\end{align*}
\]  \tag{12}

In the formula, \( \text{MAE} \) is the mean absolute error of the verification sample; \( \text{MSE} \) is the mean square deviation of the test sample; \( u_i^e \) is the expert evaluation value of the validation sample; \( u_i^{SVM} \) is the SVM calculation value of the validation sample; \( l \) is the total number of validation samples.

e) After the ideal parameters are obtained, the learning process of SVM is finished. At this time, SVM can evaluate the quality of cigarette.

At the beginning of the evaluation, only by inputting the measurement vector \( (r_{i1}, r_{i2}, \cdots, r_{im}) \) of chemical composition index of the cigarette sample \( x_i \) to be evaluated, the SVM evaluation value \( u_i^{SVM} \) of the corresponding sample can be obtained.

4.2 Example
We collected 121 groups of cigarette data from eight provinces as the original data. Each group of data is composed of 7 main chemical components of total sugar, reducing sugar, total nitrogen, nicotine, potassium oxide, chloride ion and crude protein of the tested sample and their 5 derived values, as well as the corresponding evaluation results given by the evaluation experts. Two groups of data were randomly selected from the data of each province, 16 groups in total, to form a test sample set, and the rest 105 groups of data were learning sample sets.

Since RBF kernel function has good performance under the assumption of general smoothness, it is very suitable for the case without more data and additional information. Therefore, RBF kernel function is chosen as the kernel function of SVM. We use winSVM software to solve the SVM algorithm. We divided 105 samples of the learning sample set into 7 groups, 15 samples in each group, and take Formula (13) as the goal to carry out k-fold cross validation (k=7), and the parameters of SVM are determined as: \( C=100, X=3.1 \times 10^{-3}, e^2=0.014 \). After the parameters are determined, the SVM is retrained with all 105 samples to obtain the regression function, where \( s=28, b=-0.214 \). At this time, \( \text{MAE}=1.89 \times 10^{-3}, \text{MSE}=4.87 \times 10^{-6} \). The test sample set is predicted by the established evaluation model, and the results are shown in Table 3.
Table 3. Comparison between SVM calculation of test sample and expert evaluation value

| Sample No. | Expert evaluation | SVM outputs | Relative error /% |
|------------|-------------------|-------------|-------------------|
|            | Expert evaluation | SVM outputs |                 |
|            | Sorting           | u           | Sorting           |                   |
| 1          | 71.29             | 71.59       | 12                | 0.42%             |
| 2          | 72.32             | 71.73       | 11                | -0.82%            |
| 3          | 73.03             | 74.56       | 8                 | 2.10%             |
| 4          | 70.01             | 71.14       | 14                | 1.61%             |
| 5          | 75.98             | 76.34       | 4                 | 0.47%             |
| 6          | 74.03             | 74.66       | 7                 | 0.85%             |
| 7          | 69.02             | 70.05       | 15                | 1.49%             |
| 8          | 73.60             | 74.28       | 9                 | 0.92%             |
| 9          | 79.97             | 78.65       | 2                 | -1.65%            |
| 10         | 78.01             | 79.48       | 1                 | 1.88%             |
| 11         | 72.03             | 71.56       | 13                | -0.65%            |
| 12         | 72.10             | 75.72       | 6                 | 0.83%             |
| 13         | 67.89             | 68.42       | 16                | 0.78%             |
| 14         | 74.84             | 75.85       | 5                 | 1.35%             |
| 15         | 75.06             | 74.24       | 10                | -1.09%            |
| 16         | 76.43             | 77.42       | 3                 | 1.30%             |

In the test sample set, there is a certain error between the output value of SVM and the evaluation value of experts, and the maximum relative error is 2.10%, and the minimum is 0.42%. Although there is a certain error between the output value of SVM and the evaluation value of experts, the experimental results show that the order of test set samples according to the evaluation value of experts and the output value of SVM is exactly the same. Therefore, the results of the model can be used to evaluate the quality of cigarettes, and can replace or partially replace the evaluation work of cigarette evaluation experts.

5. Conclusion

This paper proposes a method of cigarette quality evaluation based on SVM and BP neural network. The experimental results show that the method can establish a nonlinear mapping relationship between the measured values of cigarette chemical components and the evaluation values of expert quality, and reflect the preference information and reasoning mechanism of evaluation experts. This method can be used as a supplement to the expert evaluation method. It can provide a more objective and reliable method for evaluating the internal quality of tobacco after knowing the chemical composition test data of tobacco. This is of great practical significance for cigarette enterprises to carry out scientific and efficient development and design of cigarette products, reduce the workload of expert evaluation, improve work efficiency, enhance the stability of cigarette production, and thus improve market competitiveness.

It should be pointed out that with the further rationalization of the selection of learning sample set, the determination of chemical composition index and the further improvement of SVM technology, this method will be able to better evaluate the quality of cigarettes.

References

[1] Xi L. Study on Cigarette Organoleptic Quality Evaluation Based on the Behavior of Consumers[J]. Journal of Anhui Agricultural Sciences, 2011,34(2):118-134.
[2] Yang C , Hui J , Wang F , et al. Quality evaluation of chemical components in cigarette blend based on factor analysis and curve estimation[J]. Tobacco ence and Technology, 2017, 50(12):77-84.
[3] Zhang HJ, Huang QF, Wang YS. Comprehensive Evaluation of Cigarette Quality by Gray Situation Decision Analysis[J]. Guangdong chemical, 2011, 038(003):243-244,246.
[4] Shi-Qiang X , Bao-Ying W , Li-Na Z , et al. Evaluation of cigarette flavoring quality by surface-enhanced Raman spectroscopy[J]. Chemical Communications, 2018:10.1039.C8CC05689G.-
[5] Çakmakçı Karadoğan D, Önal Ö, Say Şahin D, et al. Evaluation of school teachers' sociodemographic characteristics and quality of life according to their cigarette smoking status: a cross-sectional study from the eastern Black Sea region of Turkey.[J]. Tuberk Toraks, 2017:18-24.
[6] Liu X. Study on sensory quality evaluation of cigarette smoking behavior of consumers [J]. Anhui Agricultural Science, 2011 (18): 281-283.

[7] Geiss O, Bianchi I, Barrero-Moreno J. Correlation of volatile carbonyl yields emitted by e-cigarettes with the temperature of the heating coil and the perceived sensorial quality of the generated vapours[J]. International Journal of Hygiene & Environmental Health, 2016, 219(3):268-277.

[8] Huijun Z, Yusheng W, Yujun C. Application of Principal Component Analysis in Evaluation of Cigarette Quality[J]. guangdong chemical industry, 2011,22(001):85-105.

[9] Yu D U, Xinlin W, Long Z, et al. Optimal Extraction of Licorice Pigment and Application of Its Pyrolysis Product in Quality Evaluation of Cigarette[J]. Guizhou Agricultural Sciences, 2015,76(4):28-35.

[10] Yun Y J, Dong Z T, Lai Z D, et al. Multiple Level Model in Regional Environment on Sensory Quality Evaluation,Application of Cigarette Brand[J]. southwest china journal of agricultural sciences, 2013,74(18):207-239.

[11] An-Jian Z. Application of the DTOPSIS and Grey Situation Decision in Comprehensive Evaluation of Cigarette Quality[J]. journal of anhui agricultural sciences, 2014,156(049):78-85.

[12] An-Yan X, Shan-Lian L, Mei-Zhou D, et al. The design and application of evaluation method for quality stability of cigarette production process[J]. Food & Machinery, 2017,44(38):85-103.

[13] Xiaoteng K, Haijian S U, Chunlei Z, et al. Loosening and Conditioning Process Sensory Quality Evaluation on Cigarette[J]. farm products processing, 2016,44(004):32-47.

[14] Ren ZQ, Tan HX, Pan WL, et al. Application of Principal Component Analysis in Evaluation of Cigarette Quality and its Consistence[J]. Tobacco Technology, 2013, 000(002):5-8.

[15] Jiang MH, Zheng W, Huang FY, et al. Evaluation and Verification of Inner Quality Conformity Assessment of Tobacco Material K326 in Dali Based on Cigarette Brands[J]. Hubei Agricultural Sciences, 2013, 052(013):3068-3071,3074.

[16] Ya-Xi Z. Empirical study about the comprehensive evaluation of cigarette quality based on entropy-weight TOPSIS model[J]. journal of tongling university, 2015,37(006):42-57.