Journal of Building Engineering 48 (2022) 103919

Contents lists available at ScienceDirect

Journal of Building Engineering

journal homepage: www.elsevier.com/locate/jobe

Study of thermal sensation prediction model based on support vector classification (SVC) algorithm with data preprocessing

Tingzhang Liu *, Linyi Jin, Chujun Zhong, Fan Xue

School of Mechatronic Engineering and Automation, Shanghai University, China

ARTICLE INFO

Keywords:
Thermal sensation
Data cleaning
Edited nearest neighbour
Synthetic minority oversampling technique
Support vector classification

ABSTRACT

In order to meet people's demand for comfort, indoor thermal environment often needs to be adjusted. Nevertheless, HVAC target parameters are often over-setted due to cognitive bias, resulting in an uncomfortable environment and waste of energy. Therefore, a thermal sensation prediction model is required to advise the setting of environmental and individual parameters. An accurate mode for predicting thermal sensation encountered severe challenges because of sensor errors, environmental noise, subjective thermal sensation differences, and sample data imbalance. The concrete aim of this paper is to verify the potential of using Support Vector Classification (SVC) algorithm to predict thermal sensation vote (TSV) based on Edited Nearest Neighbour (ENN) and Synthetic Minority Oversampling Technique (SMOTE), named combined ENN + SMOTE + SVC method. Firstly, for the problem of outliers in the dataset, the ENN method was adopted to clean the raw data. Secondly, SMOTE method is used to expand the training data which has the problem of sample imbalance. Finally, SVC algorithm is adopted to build a thermal sensation prediction model. The results show that the model built by the combined ENN + SMOTE + SVC method can achieve better performance than PMV index and other classic classification algorithms.

1. Introduction

According to surveys, people spend up to 80% of their time indoors [1]. Indoor conditions have therefore far-reaching implications for their health, general well-being and performance. A literature of indoor environmental conditions found that people tend to pay more attention to thermal comfort than visual, auditory, and indoor air quality comfort [2]. Heating, ventilation and air conditioning (HVAC) systems can directly change the indoor thermal environment and meet people's demands for thermal comfort. Air conditioning systems in residential and office buildings account for about 20–30% of the world's energy consumption [3,4]. People generally have deviations in their perception of the comfortable temperature range, that is, the indoor set temperature is often too high or too low, which leads to an uncomfortable environment and waste of energy. Therefore, a thermal sensation prediction model is required to advise the setting of the indoor temperature, or to adjust other environmental or individual parameters when the indoor temperature is unmodifiable, so as to make the indoor environment comfortable. The thermal sensation obtained by the model can be used as a constraint to find the optimal HVAC settings and other parameters to achieve energy saving.

Generally speaking, there are two kinds of models for predicting thermal sensation – heat balance models and adaptive models. The heat balance model assumes that the thermoregulatory system of human could maintain a basically constant internal body

* Corresponding author.
E-mail address: liutzhcom@oa.shu.edu.cn (T. Liu).

https://doi.org/10.1016/j.jobe.2021.103919
Received 7 June 2021; Received in revised form 17 November 2021; Accepted 18 December 2021
Available online 22 December 2021

temperature. In order to maintain this temperature, human body will take physiological reactions, i.e., modification of skin temperature or sweat secretion, to adapt any environments that may cause thermal imbalance [5]. Fanger developed the Predicted Mean Vote (PMV) index, which mathematically predicts the average ASHRAE thermal sensation vote (3 (hot); 2 (warm); 1 (slightly warm); 0 (neutral); −1 (slightly cool); −2 (cool); and −3 (cold)) of a large group of individuals [6–8]. Since the PMV represents a mean vote, which ignore the individual votes. Fanger subsequently developed the Predicted Percentage of Dissatisfied (PPD) index, which is correlated with the PMV index [6].

The heat balance model is not suitable for everyday work and living environments, so adaptive model is designed for those environments. The theory of adaptive model holds that people will interact and adapt to their environment to restore their thermal comfort [9,10]. Adaptation can be divided into three categories — physical, behavioural and psychological adaptations [11]. Numerous thermal comfort theories related to the adaptive model have been put forward. The theoretical adaptive models, called aPMV [1]and nPMV [12], which considered multiple factors such as culture, climate, social, psychological and behavioural adaptations, are proposed respectively. The view 'a person's reaction to a temperature which is less than perfect will depend very much on his expectations, personality and what else he is doing at the time' shows the role of expectation in thermal comfort [13]. Compared to people with low degree of control in the same environment, those with high control over the environment are more likely to be satisfied [14].

With the rapid development of information technology (especially artificial intelligence technology), researchers have discovered the advantages of artificial intelligence algorithms and applied them for thermal sensation prediction. Backpropagation neural network technology has been used to construct a personal thermal sensation model, which divides personal thermal sensation into three conditions: high temperature, neutral and cold [15]. Comparing to the ASHRAE 7-scale sensations, this model outputs the thermal sensations in less detail. Based on the database of the RP-884 adaptive model project, Grabe studied the potential of artificial neural networks to predict thermal sensation votes (TSV) [7]. Although its performance is superior to the classic PMV index in terms of prediction quality, its prediction accuracy (55% for training set and 50% for cross-validation set) is generally not good enough. This method improves the model just by feature selection, adjusting the size of training and test sets, and using regularization, but without preprocessing the complex data itself.

Li et al. developed an estimation model based on the Takagi–Sugeno (TS) fuzzy model considering the difference in human thermal sensation under different activities [16]. Although the proposed model can estimate human thermal sensation under unstable environmental conditions, this experiment has obvious narrowness no matter in the age, physical condition, activity state, or experimental environment, and the experimental results cannot be widely applied. Salehi, Ghanbaran and Maerefat examined the effectiveness of six different intelligent approaches for predicting thermal sensation and demand using occupants' physiological factors. Regarding thermal demand, it was found that the accuracies of the Gaussian Process Regression (GPR) and PMV models were 86% and 69% [17]. The support vector machine (SVM) algorithm has been utilised in the thermal comfort research area. Megri, Naqa and Haghighat applied the support vector regression (SVR) to develop the thermal sensation model [18]. It is claimed that their research showed the potential of using the SVM to generate the thermal index of specific populations. Lai Jiang et al. developed a new personal thermal modelling method based on C-Support Vector Classification (C-SVC) algorithm, which can be integrated into a personalized conditioning system to optimize its operation and control [19]. As one of the widely used classification algorithms, SVC can avoid "dimensional disaster". When the parameters and kernel functions are selected properly, its generalization ability is greatly improved.

In the data collection, sensor errors, environmental interference and other factors cause the data to contain abnormal points. The criterion for determining abnormal points is whether they are outliers. Outliers do not meet the system characteristics contained in most data, which affects the generalization ability of the model and the reliability of the data. A high-precision thermal sensation prediction model needs to be built after eliminating potential disturbance. However, the effect of data cleaning in thermal prediction research has not been tested at present. Moreover, the number of samples between the classifications is extremely imbalanced, so the model cannot be effectively trained, which seriously affects its generalization ability. Most researches only focus on data partitioning or improving modelling algorithms when facing the same low training accuracy and verification accuracy, but fail to solve the problem essentially.

SVM is one of the popular machine learning algorithm, many improvements to the above problems have been proposed. Jayadeva and Chandra formulated SVM in a way that two proximal class hyperplanes are derived termed as TWSVM, theoretically reduces the training cost to $1/4^{th}$ of SVM [20]. Improvements on TWSVM had been carried out by introducing the regularization terms into the objective functions of the twin bounded support vector machine (TBSVM) [21]. Recently, Fan et al. proposed entropy-based fuzzy support vector machine (EFSVM) for imbalance learning [22]. Borah and Gupta proposed a robust twin bounded support vector machine algorithm based on truncated loss functions to improve the generalization performance of TBSVM with reduced sensitivity to noise and outliers and to handle class imbalance learning as well [23]. These methods can further optimize SVM, and its effect on thermal comfort prediction remains to be tested.

This paper mainly studies the potential of the combined ENN + SMOTE + SVC method to improve the accuracy of thermal comfort prediction. Field experiment data were used to train the model to output thermal sensation votes. Firstly, for the problem of outliers in the data, the ENN method was adopted to clean the raw data. Secondly, the cleaned data still has the problem of sample imbalance, and SMOTE method is used to expand the training data. Finally, the "Grid Search" method is used to find the optimal SVC parameter C and kernel parameters γ, and SVC algorithm is adopted to build a thermal sensation prediction model.

The rest of the paper is arranged as follows: section 2 provides a brief introduction to three methods: SVC, ENN, SMOTE. Section 3 gives a brief overview of the modelling methods, datasets, and evaluation indicators. Section 4 performs feature selection and data processing on the data used in this paper, then uses "grid search" to find the optimal penalty parameter C and the kernel parameter γ, and finally obtains a thermal sensation prediction model with high generalization ability. Section 5 introduces different algorithms for

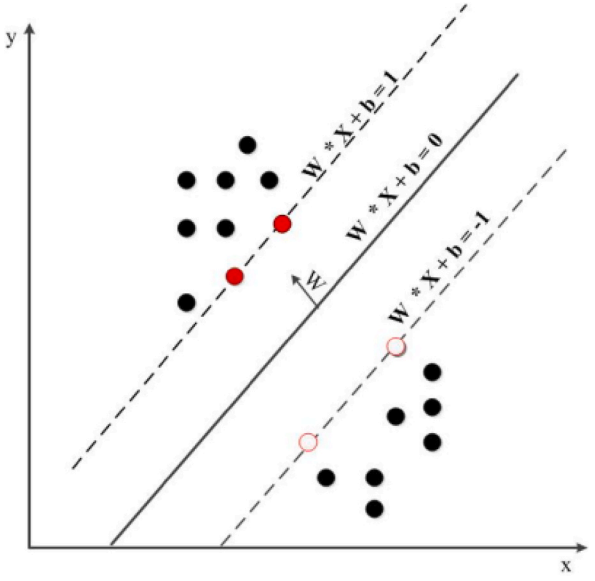


Fig. 1. The decision boundary, interval boundary, and support vector.

comparison, highlighting the superiority of this method. Section 6 draws a conclusion.

2. Methodology

2.1. SVC algorithm

The Support Vector Machine (SVM) is a supervised learning algorithm that improves the versatility of machine learning by minimizing structured risk, and has been widely used in the classification field. The Support vector classification (SVC) is a two-class model whose learning strategy is to maximize the interval and eventually transform it into a solution of convex quadratic programming [24]. Assuming the total number of dataset is N, arrange the collected data as input-output pairs, which can be expressed as (x_i, y_i), i = 1,2, …,N. x_i is an input vector that includes factors such as the environment and the individual; y_i is an output variable that includes only the individual's thermal sensation in the environment. Suppose $y_i = 1$, a positive class, which is the first type of thermal sensation; $y_i = -1$, a negative class, which is another type of thermal sensation.

The idea of SVC is to use the "hyperplane" as the decision boundary (see Eq. (1)), to separate the learning targets into positive and negative classes, and make the point-to-plane distance (see Eq. (2)) of any sample greater than or equal to 1.

$$w^T x + b = 0 \tag{1}$$

$$y_i\left(w^T x_i + b\right) \geq 1 \tag{2}$$

The parameters w and b are the normal vector and intercept of the hyperplane, respectively. The decision boundary that satisfies this condition actually constructs two parallel hyperplanes as the interval boundary to discriminate the classification of the sample:

$$w^T x_i + b \geq +1, \Rightarrow y_i = +1 \tag{3}$$

$$w^T x_i + b \leq -1, \Rightarrow y_i = -1 \tag{4}$$

The samples above the upper interval boundary belong to the positive class (Eq. (3)), and the samples below the lower interval boundary belong to the negative class (Eq. (4)). The distance between two interval boundaries is defined as the margin $\delta = \frac{2}{w}$, and the positive and negative samples located on the interval boundary are support vectors. Fig. 1 shows the decision boundary, interval boundary, and support vector, respectively.

To find the maximum interval to divide the hyperplane, that is, the constraint parameters w and b that satisfy Equation (5) and Equation (6) must be found to maximize δ:

$$\underset{w,b}{max} \frac{2}{\|w\|} \tag{5}$$

$$s.t. y_i\left(w^T x_i + b\right) \geq 1, i = 1, 2, \cdots, N. \tag{6}$$

For nonlinear separable problems, hypersurfaces exist in the feature space to separate positive and negative classes. Radial basis kernels can be used to map the original feature space to a more infinite dimensional space to solve this type of problem. In addition, the

radial basis kernel function (Eq. (7)) is used to avoid the display calculation of the inner product.

$$K(x_i, x_j) = \exp\left(-\frac{\left\|X_i - X_j^2\right\|}{2\sigma^2}\right) = \exp\left(-\gamma\left\|X_i - X_j^2\right\|\right) \tag{7}$$

σ is radial basis kernel bandwidth. x_i and x_j are two different input vectors in the feature space. $\gamma = 1/2\sigma^2$ is an RBF-specific parameter that is used to optimize model performance.

In practical tasks, it is difficult to determine the appropriate kernel function to make the training samples linearly separable in the feature space. Even if this kernel function is found exactly, it is difficult to conclude that the result is not caused by overfitting. One way to mitigate this problem is to allow a few errors in the calculation of the separation. In other words, slack variable ξ and corresponding penalty factor C are introduced to allow certain points to be on the wrong side of the interval surface. According to the dual transformation and the introduced variables, Equation (5) can be rewritten as Equation (8):

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{8}$$

For multi-classification problems, “one-against-all” strategy, which m classifications establish m decision boundaries, and each decision boundary can judge against one class, is applied to solve this type of problem. This paper needs to classify 7 levels of thermal sensation, so 7 decision boundaries are needed to correctly classify each class.

The model established by SVC only relies on a few samples, so that the model is susceptible to noise. It is necessary to clean the data before training the model.

2.2. ENN method

The Edited Nearest Neighbour (ENN) method is a data cleaning method, which mainly applies the nearest neighbour rule to find and remove those samples that are not friendly to neighbours [25]. There are two forms: "half" and "all". In the "half" form, if more than 50% of the neighbour samples for a sample belong to the same class, the sample will be saved. In the "all" form, 100% is required. This paper uses “all” form to thoroughly clean the data.

Except for the one class with the smallest number of samples, the rest were cleaned. For sample j, the feature vector is x_j , the specific process is as follows:

(1) Apply the nearest neighbour rule to find its k neighbours, denoted as x_j (near), and near∈{1, …, k};
(2) Determine whether all (or half) of k nearest neighbours belong to the same class as sample j. Yes, sample j is retained; otherwise, sample j is discarded.

The number of samples is greatly reduced after the data cleaning. However, the data set used in this paper has the problem of sample imbalance (see details in Table 5), data cleaning may cause it to intensify. Because of this, the model cannot be effectively trained, which seriously affects the generalization ability of the model. Therefore, it is necessary to expand the training set so that the model can be fully trained.

2.3. SMOTE method

When faced with imbalanced dataset, standard classification algorithms tend to predict only the classes that make up most of the data. A small number of samples are considered noise and are usually ignored. This can seriously affect the generalization ability of the model. The oversampling strategy is one of the basic methods to solve the problem of class imbalance.

The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method in which the minority class is oversampled by creating “synthetic” examples rather than by replacement [26]; that is, it is based on “interpolation” to synthesize new samples for the minority classes. The algorithm borrows KNN technology, and the specific steps for generating new samples are as follows:

(1) For each sample x in the minority class, calculate its distance to all samples in the minority class sample set S_{min} using the Euclidean distance as the standard, and obtain its k nearest neighbours;
(2) Set a ratio according to the sample imbalance ratio to determine the sampling magnification K. For each minority class sample x, randomly select several samples from its k nearest neighbours, assuming the selected neighbour is x_n;
(3) For each randomly selected neighbour x_n, a new sample is created with the original sample according to the following Equation (Eq. (9)):

$$x_{new} = x + \theta \times |x - x_n| \tag{9}$$

θ is a random factor, θ ∈ (0,1).

3. Dataset and modelling methods

3.1. Preparation of the dataset

This paper applied the RP-884 Adaptive Model Project dataset [27] for its advantages of thermal comfort field data from all over the world, standard calibration and sufficient parameter collection. A classification method for modelling is applied in this paper, so the value of thermal sensation vote needs to be discrete. Therefore, the dataset with thermal sensation as the standard 7-scale is selected.

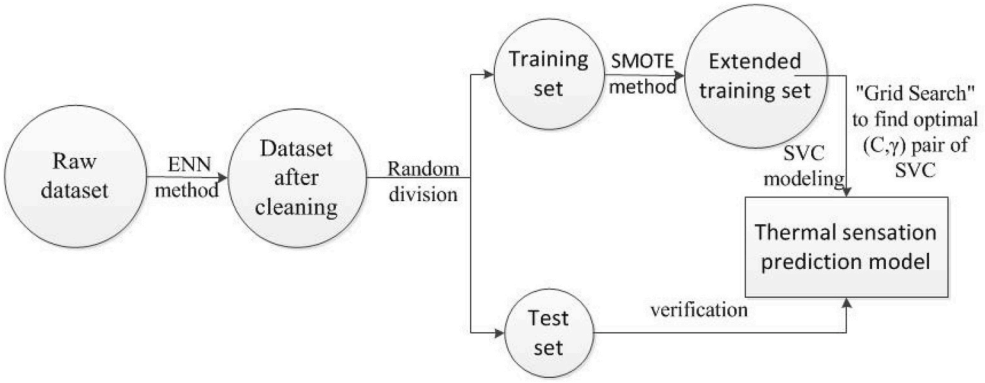


Fig. 2. The overall framework of the method.

Its locations include the United Kingdom, Canada, Pakistan, Athens, Australia and other countries, covering the climate of the ocean coast, Mediterranean, subtropics and desert. The building is dominated by office areas. If some important values (such as indoor temperature, thermal sensation vote, etc.) are missing, the sample is deleted. After preliminary processing of the data, the total number of datasets was reduced from 8001 to 7222. After preliminary processing of the data set, there are still outliers and imbalances in the data, which will be further introduced in Section 4.2 and Section 4.3. The number of samples for each scale can be seen in Table 5. In this paper, the dataset is randomly divided into a training set (TDS, 80%) and a test set (TestDS, 20%). The training set is used to train the model, and the test set is to verify the generalization ability of the model.

3.2. Modelling method

The combined ENN + SMOTE + SVC method is applied to build thermal sensation prediction model in this paper. Firstly, for the problem of outliers in the data, the ENN method was adopted to clean the raw data. Secondly, the cleaned data still has the problem of sample imbalance, so SMOTE method is used to expand the training data. Finally, the "Grid Search" method is used to find the optimal (C, γ) pair, and SVC algorithm is adopted to build a thermal sensation prediction model. The overall framework of the method is shown in Fig. 2.

3.3. Evaluation indicators

In order to verify the performance of the proposed ENN + SMOTE + SVC model, several reasonable evaluation indicators need to be designed. The most commonly used indicator for evaluating a model is accuracy. F1 score, ROC curve and AUC are added as evaluation indicators to comprehensively evaluate model performance.

(1) Accuracy rate

Accuracy is defined as the ratio of correctly classified samples to the total sample, the higher the accuracy, the more accurate the model is to a certain extent. Its calculation equation is:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

Among them, TP is a true positive; FP is a false positive; FN is a false negative; TN is a true negative.

(2) F1 score

F1 score is a harmonic average of precision and recall. Recall and precision are a pair of contradictory measures. Generally speaking, when the precision is high, the recall is often low. FI value can comprehensively evaluate the precision and recall of the model. The closer the F1 score is to 1, the better the generalization ability of the model. The calculation equation is:

$$F1 = 2^* \frac{precision^* recall}{precision + recall} \tag{11}$$

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

(3) ROC and AUC

The ROC is called the receiver operating characteristic curve. It is a graph constructed with true positive rate (TPR) as the ordinate

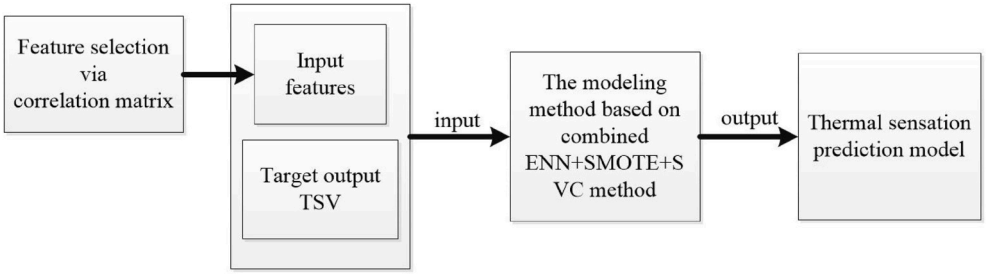


Fig. 3. The training process of the thermal sensation prediction model.

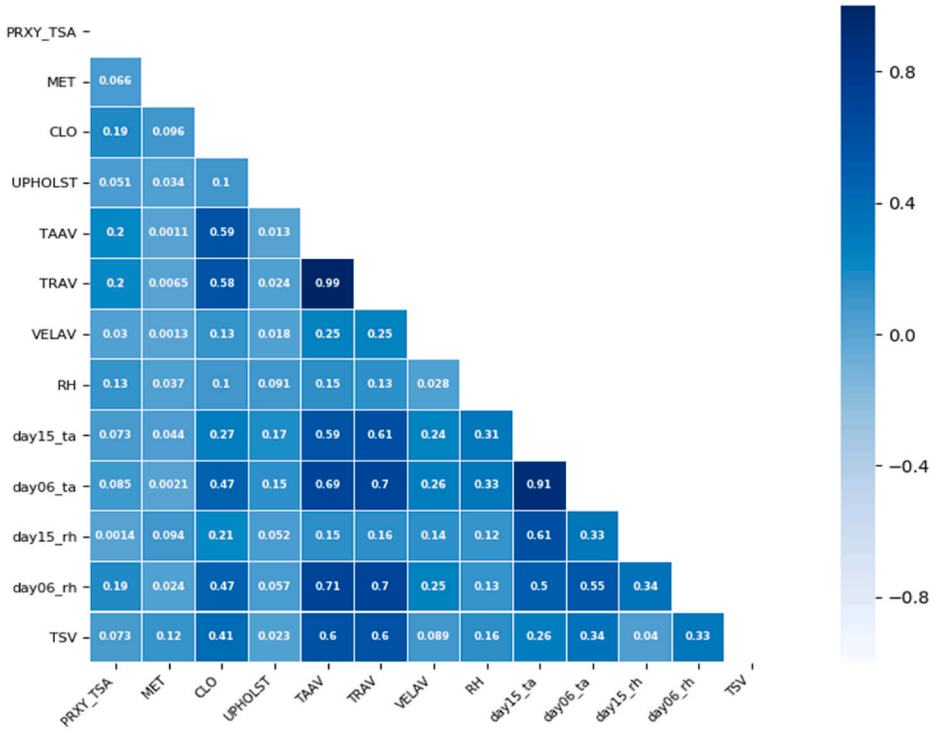


Fig. 4. Correlation matrix of 12 features and TSV.

and false positive rate (FPR) as the abscissa. This curve shows the balance between coverage and accuracy. Ideally, TPR should be close to 1, and FPR should be close to 0.

The AUC value is the area under the ROC curve, which reflects the ability of the classifier to sort the samples. The AUC value is not sensitive to whether the sample classes are balanced and is often used to evaluate the performance of classifiers. The larger the AUC value, the better the performance of the classifier.

4. Results

First, the optimized input variables should be selected through method of feature selection; then all data should be arranged as input and target output pairs to fit the combined ENN + SMOTE + SVC method. The training process of the thermal sensation prediction model is shown in Fig. 3.

Experiments are conducted on python3.5 with the widely used scikit-learn and imbalanced-learn.

4.1. Feature selection

Regarding feature selection, many researches have been conducted through expert knowledge or experiments. Lai Jiang et al. directly used the six parameters of the PMV index of Fanger as input variables [19]; Jörn von Grabe et al. combined Fanger's PMV index and related experimental research to select input variables [7]. Correlation analysis is a common method to analyse the influence of variables on each other. This paper implements feature selection based on the correlation matrix of each feature and TSV. The theoretical basis used by the correlation matrix is covariance. The larger the covariance value between the two features, the more

Table 1
Ranking of 12 features based on coefficient with TSV.

Rank	Feature name	Correlation coefficient with TSV
1	TAAV	0.6
2	TRAV	0.6
3	CLO	0.41
4	day06_ta	0.34
5	day06_rh	0.33
6	day15_ta	0.26
7	RH	0.16
8	MET	0.12
9	VELAV	0.089
10	PRXY_TSA	0.073
11	day15_rh	0.04
12	UPHOLST	0.023

Table 2
Model performance with different input features.

Input features	Training accuracy	Test accuracy	F1 score	AUC
TAAV, TRAV, CLO, day06_ta, day06_rh, day15_ta	0.571	0.514	0.51	0.72
TAAV, TRAV, CLO, MET, VELAV, RH	0.544	0.502	0.50	0.71

Table 3
Effect of feature number on SVC model's generalization ability.

Number of features	12	11	10	9	8	7	6
Training accuracy	0.723	0.725	0.680	0.690	0.696	0.586	0.571
Test accuracy	0.575	0.579	0.541	0.507	0.514	0.520	0.514
F1 score	0.57	0.58	0.54	0.51	0.51	0.52	0.51
AUC	0.75	0.75	0.73	0.71	0.72	0.72	0.72

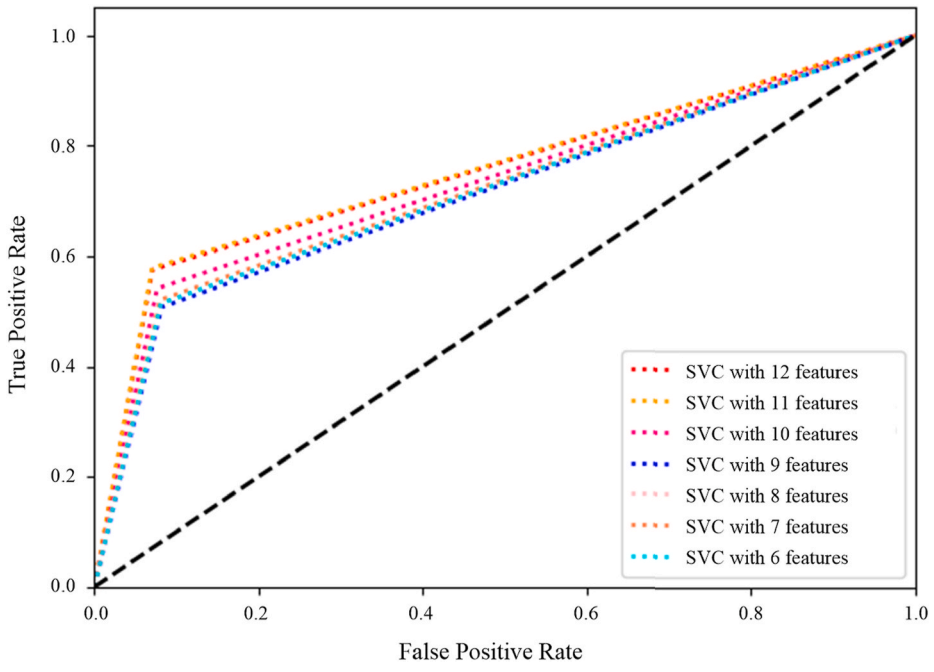


Fig. 5. ROC curves with different number of features.

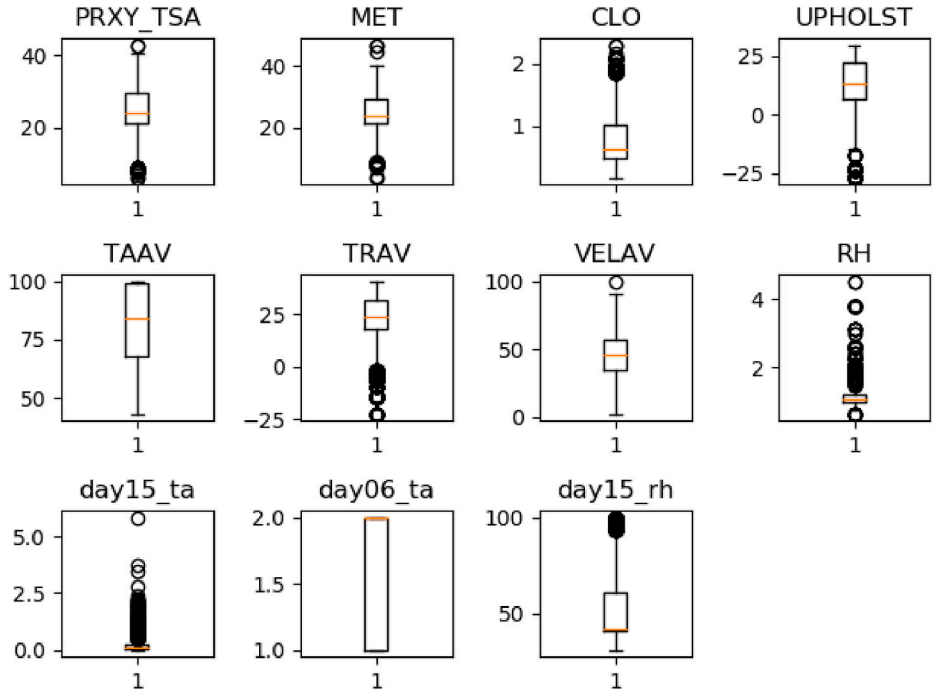


Fig. 6. Boxplot of 11 input features.

related they are. Each feature is ranked according to the correlation coefficient with TSV, providing a theoretical basis for feature selection.

In the public datasets, each dataset may come from different field surveys, so the same features will be few. We only found 12 common and independent variables, which are: average of three heights' air temperature (TAAV), outdoor 6 a.m. (min) air temperature on day of survey (day06_ta), outdoor 3 p.m. (max) air temperature on day of survey (day15_ta), average of three heights' mean radiant temperature (TRAV), ensemble clothing insulation (CLO), insulation of the subject's chair (UPHOLST), relative humidity (RH), outdoor 6 a.m. (max) relative humidity on day of survey (day06_rh), outdoor 3 p.m. (min) relative humid on day of survey (day15_rh), average metabolic rate of subject (MET), average of three heights' air speed (VELAV), thermal acceptability defined as $-1.5 \leq$ ASHRAE thermal sensation scale (ASH)≤ 1.5 (PRXY_TSA). The correlation between these 12 features and the TSV are mainly analysed. Through correlation analysis and experimental demonstration, the best feature combination can be obtained as an input variable. Fig. 4 shows the correlation matrix of 12 features and TSV. According to the magnitude of the correlation with TSV, each feature can be accurately ranked, as shown in Table 1. From the ranking in the table, the top 6 features with correlation coefficients are TAAV, TRAV, CLO, day06_ta, day06_rh, day15_ta, which are inconsistent with the 6 features selected by the PMV index (TAAV, TRAV, CLO, MET, VELAV, RH).

The above SVC model id used to predict thermal sensation votes, the parameter C of SVC is 1.0 and kernel parameter γ is 0.1. The selection of parameters will be introduced in Section 4.4. Taking these two groups of 6 variables as the input variables of the model, Table 2 shows the performance of the model. It can be seen that when the top 6 features as input variables, the test accuracy is 0.514, F1 score is 0.51, and AUC is 0.72. While using the 6 parameters of the PMV index as input, the test accuracy is 0.502 (-0.012), F1 score is 0.50 (-0.01), and AUC is 0.71 (-0.01). The performance of the two models is similar (the top 6 feature model is slightly better), which shows that compared with the traditional feature selection method, the feature selection based on the correlation matrix in this paper is effective.

In order to study the effect of the number of selected features on the generalization ability of the model, the first several features according to the rank of the correlation coefficient are successively selected as input variables. The results are shown in Table 3. It can be concluded from Table 3 that when the inputs are the first 11 features, the test accuracy rate, F1 score and AUC are the highest, and the model has the best generalization ability. Fig. 5 also illustrates this because under the same false positive rate, the true positive rate of first 11 features is higher than the other curves.

Therefore, the first 11 features were selected as input variables through the correlation matrix selection, including TAAV, TRAV, CLO, day06_ta, day06_rh, day15_ta, RH, MET, VELAV, PRXY_TSA, day15_rh, to optimize the model's generalization ability. Although the generalization ability of the model has been effectively improved after feature selection, it still needs to be further improved. The highest training accuracy rate is 0.725, and the test accuracy rate is even lower, which indicates that the model has not been effectively trained under the condition of the enough amount of data. This may be caused by outliers in the data. To further improve the generalization ability of the model, outliers should be cleaned.

Table 4
Generalization ability of the model with different operations.

Methods	Operations		Sample number	Training accuracy	Test accuracy	F1 score	AUC
SVC	Data without cleaning		7222	0.725	0.579	0.58	0.75
SVC + ENN	Data cleaning using ENN	"half" form	4067	0.912	0.844	0.84	0.91
		"all" form	1907	0.971	0.895	0.90	0.94
SVC + ENN + SMOTE	Data cleaning "all" form and sample equalization		5572	0.991	0.922	0.92	0.96

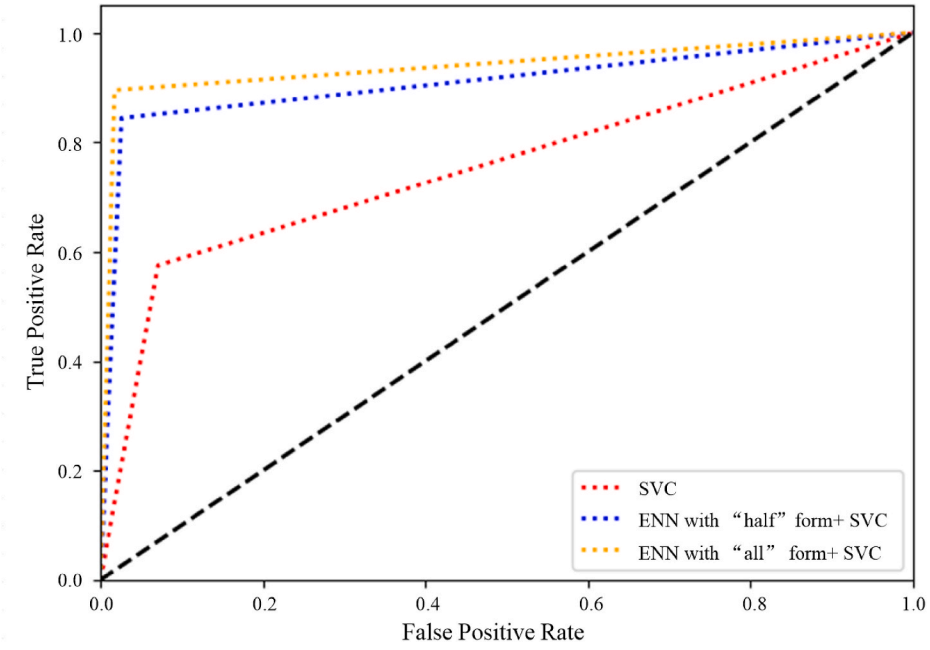


Fig. 7. ROC curve for data cleaning vs. data not cleaning.

Table 5
Number of samples in each class before and after data cleaning.

Class	Before data cleaning	Proportion of total sample size	After data cleaning	Proportion of total sample size
Total	7222	1	1907	1
-3	475	6.58%	151	7.92%
-2	532	7.37%	62	3.25%
-1	905	12.53%	61	3.20%
0	2952	40.88%	995	52.18%
1	1385	19.18%	235	12.32%
2	699	9.68%	129	6.76%
3	274	3.80%	274	14.37%

4.2. Data cleaning

When the number of samples is large enough, the training accuracy and test accuracy are still low, it may be caused by abnormal data. The boxplot is used as data analysis to prove this. Boxplot is used to reflect the characteristics of the original data distribution, and points that are not between the lower and upper edge lines are considered outliers. Fig. 6 is a boxplot of each feature. From Fig. 6 it can be easily seen that except for TAAV, VELAV, day06_ta, there are many outliers in the other features. Outliers are points that are significantly different from other data objects. Most data mining methods treat outliers as noise or anomalies and discard them [28]. Although this may lead to the loss of extreme cases, it can greatly improve the performance of the model (for some special models that rare events are particularly important, anomaly mining can be used to optimize). Therefore, cleaning data is required before modelling.

There are two forms of the ENN algorithm: "half" and "all". Contrast experiments are conducted to study which form of ENN algorithm is best for the generalization ability of the model.

It can be seen from Table 4 that the greater the degree of data cleaning, the better the generalization ability obtained by the model, but there may be excessive cleaning. From the ROC curve (Fig. 7), under the same false positive rate, the true positive rate of the data

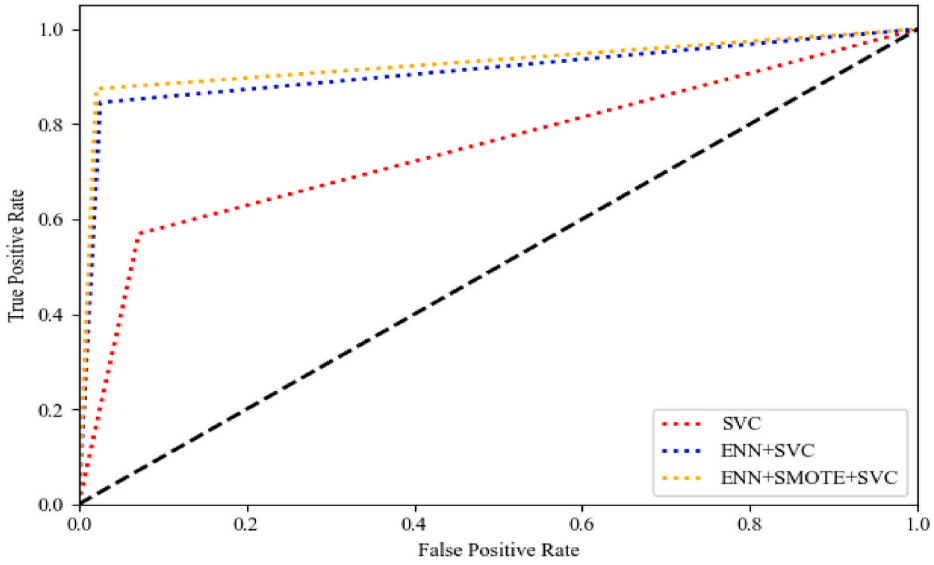


Fig. 8. ROC curve after different data processing procedures.

cleaning of the "all" form of the ENN algorithm is higher than that of the unwashed and the "half" form's. Therefore, using the "all" form of the ENN algorithm for data cleaning, the generalization ability of the model has improved dramatically:

the training accuracy rate is 0.971 (+0.246), the test accuracy rate is 0.895 (+0.316), and the F1 score is 0.90 (+0.32) and AUC of 0.94 (+0.19). However, because the data retention rules were too harsh, the number of samples decreased from the original 7222 to 1907. While using the "half" form of the ENN algorithm for data cleaning, although the improvement effect is not as good as in the "all" form (training accuracy rate is 0.912(+0.187), test accuracy rate is 0.8444(+0.265), F1 score is 0.84(+0.26), and AUC is 0.91(+0.16)), the number of retained samples is more than that in the "all" form. The total number of outliers obtained from the boxplot analysis is 3932, and the number of data cleaned by "half" form (3155) is near to that of outliers. This shows that the ENN algorithm in "half" mode cleans most of the outliers, which can improve the model performance.

Since the number of samples processed in this paper is sufficient, and as mentioned above, the model under the "all" mode cleaning has better generalization ability, this paper uses the "all" mode for data cleaning. However, although the model's generalization ability has been greatly improved, the problem of sample imbalance still exists and has even been exacerbated. Table 5 shows the number of samples in each class before and after data cleaning, samples are classified according to TSV value. Before the data cleaning, 40.8% of the samples had a TSV value of 0; after the data cleaning, the proportion increased to 52%. The next section will study the impact of the sample imbalance problem on the model's generalization ability.

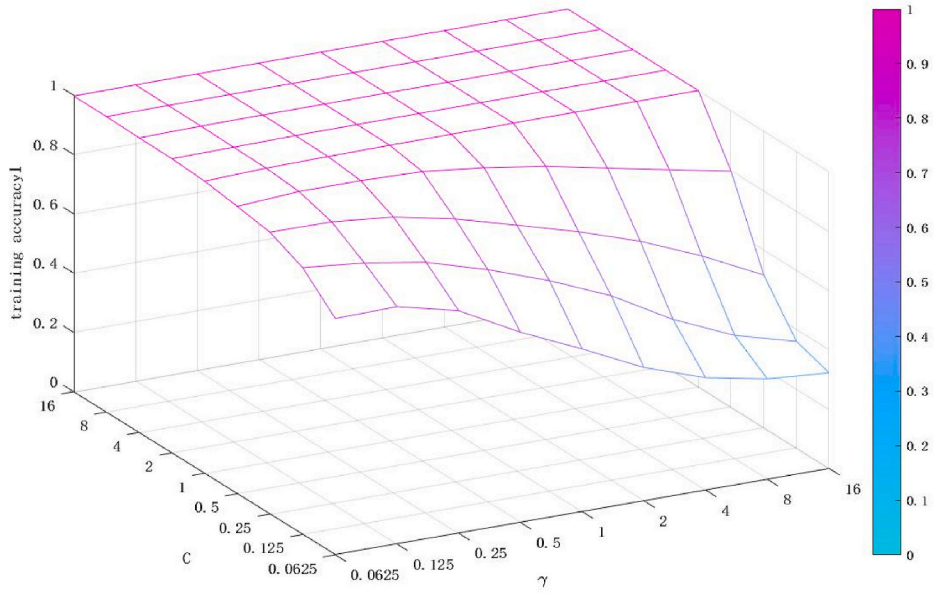
4.3. Sample equalization

The thermal sensation is divided by 7-scale (3 (hot); 2 (warm); 1 (slightly warm); 0 (neutral); −1 (slightly cool); −2 (cool); and −3 (cold)). However, because people stay indoors most of the time, extremely cold and hot conditions rarely occur, so it is normal that the collected data has the problem of sample imbalance. Table 5 shows the number of each class before and after data cleaning, before data cleaning the number of samples in class of 0 is 2952, while the number of samples in class of −2 are 532, and the difference is more than five times, which confirms the problem of sample imbalance. After data cleaning the number of samples in class of 0 is 995, while the number of samples in class of −1 and class of −2 are around 60, and the difference is more than ten times, indicating that data cleaning aggravated the problem of sample imbalance. The problem of sample imbalance may affect the generalization effect of the model. Therefore, the SMOTE algorithm is introduced to the training set to achieve sample equalization and to study whether the generalization ability of the model can be further improved.

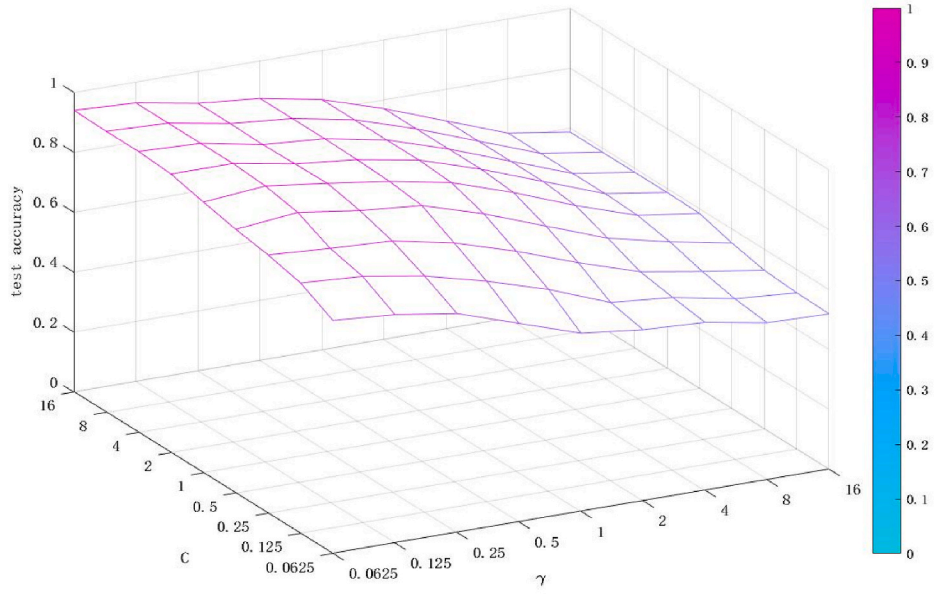
From the ROC curve (Fig. 8), it can be seen that under the same false positive rate condition, after using the SMOTE algorithm on the training set to expand the sample, the trained model will get the largest true positive rate, which shows that solving the problem of sample imbalance can improve the model generalization. As can be seen in Table 4, after the SMOTE algorithm training accuracy reaches 99.1% (+2%), test accuracy reaches 92.2(+2.7%), F1 score reaches 0.92(+0.02), and AUC reaches 0.96(+0.02). Sample equalization using the SMOTE algorithm can further improve the generalization ability of the model.

4.4. Parameter adjustment

In the SVC model, there are two critical parameters, the penalty parameter C and the kernel parameter γ. In practical applications, their values largely determine the performance of the SVC [29]. The parameter C is the penalty coefficient, which indicates the tolerance of the error. Higher C value indicates that the error cannot be tolerated and it is easier to overfit; smaller C value is easier to underfit. Therefore, too large or too small C value may make the model's generalization ability worse. The parameter γ comes with the RBF function as a kernel function, which implicitly determines the distribution of the data after it is mapped to the new feature space.



(a) training accuracy



(b) test accuracy

Fig. 9. Influence of different (C, γ) pairs on training and test accuracy.

The larger γ value, the fewer support vectors, the better the training result, but the more likely it is to cause overfitting; and the smaller γ value, the more support vectors, it is easier to cause underfitting.

In this research, the parameters C and γ have been optimally selected by the "Grid Search" method [30]. A series of C and γ values are calculated from 2^A, and then all possible combinations of (C, γ) pairs are generated, where A ranges from (−4, −3, −2, −1, 0, 1, 2, 3, 4). Each step a pair (C, γ) is automatically selected and then applied to the training model. If there are multiple (C, γ) pairs for best performance, we prefer to choose smaller (C, γ) pairs to prevent overfitting and to apply to other datasets.

From the previous research, the changes in F1 score and AUC are consistent with the changes in test accuracy, so here we focus on the influence of the changes of (C, γ) pair on the training accuracy and test accuracy. It can be seen from Fig. 9 that different (C, γ) pairs have a great impact on the generalization ability of the model. From Fig. 9a, when C≥1, no matter how the value of γ changes, the training accuracy rate has reached the maximum value and will not change again. This shows that when C is large enough, the model is

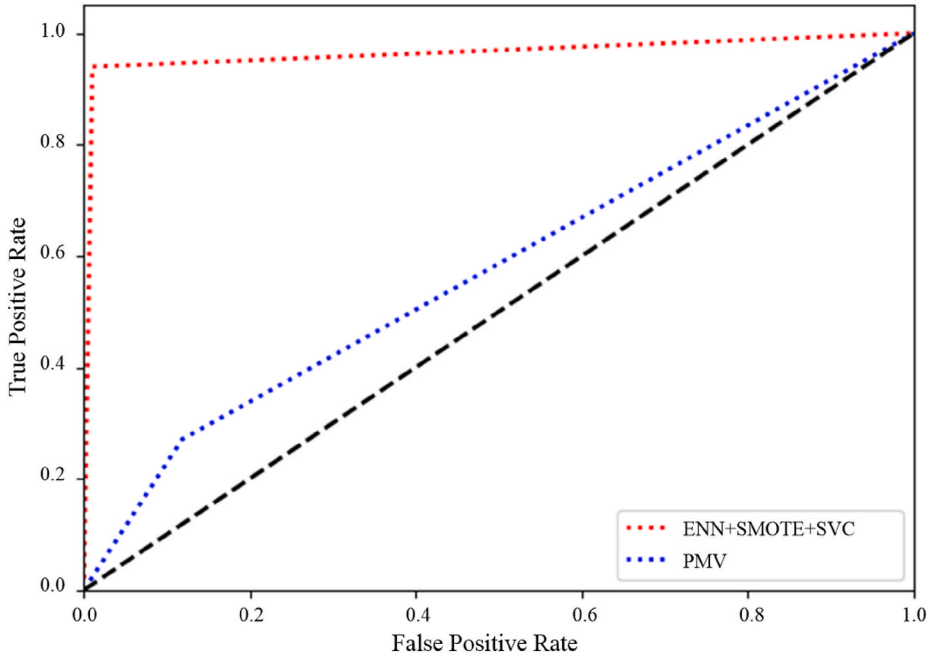


Fig. 10. ROC curves of ENN + SMOTE + SVC method vs. PMV index.

Table 6
Effects of different algorithms on model performance.

Algorithm	Training accuracy	Test accuracy	F1 score	AUC
PMV	0.284	0.271	0.27	0.58
ENN + SMOTE + AdaBoost	0.546	0.539	0.54	0.73
ENN + SMOTE + Gauss Bayes	0.633	0.503	0.50	0.71
TBSVM	0.74	0.538	0.70	0.79
ENN + SMOTE + KNN	0.977	0.859	0.86	0.92
ENN + SMOTE + CART	1.0	0.866	0.86	0.92
ENN + SMOTE + SVC	0.994	0.940	0.94	0.96

trained better, but overfitting may occur. From Fig. 9b, when the C value remains unchanged, the larger the γ value is, the smaller the test accuracy is, and the worse the generalization ability of the model is; when the γ value remains unchanged, the larger the C value, the greater the test accuracy and the stronger the generalization ability of the model. If $\gamma = 0.0625$, when $C >= 1$, the training accuracy reaches the maximum (1.0), and when $C >= 4$, the test accuracy reaches the maximum (0.940). Therefore, $C = 4$ and $\gamma = 0.0625$ are chosen as the optimal parameters of the SVC algorithm. Proper (C, γ) pair further improves the generalization ability of the model.

This section studies the influence of different features, data processing and modelling methods on the thermal sensation prediction model. Through the correlation analysis of the variables in the data set, the top 11 variables with the highest correlation with TSV are selected as input features; The ENN method is used to clean the raw data containing outliers. After cleaning with the "all" mode, the amount of samples is reduced from the original 7222 to 1907, and the accuracy of the verification set of the prediction model established by SVC is increased from 57.9% to 89.5%; In order to solve the problem of sample imbalance in the data set, SMOTE is used to expand the sample, and the accuracy of the model validation set is further adjusted to 92.2%; finally, the grid search method is used to find the optimal parameter pair ($C = 4$, $\gamma = 0.0625$), the accuracy of the prediction model is increased to 94.0%, the FI score is 0.94, and the AUC is 0.96.

5. Discussions

The previous sections have proved that the combined ENN + SMOTE + SVC method proposed in this paper is a powerful tool for predicting thermal sensation. For a given condition, the combined ENN + SMOTE + SVC method is based on empirical data rather than lab data, so it is more consistent with field observations than the PMV; Sometimes using PMV prediction will get ridiculous results (such as −5, 5, etc.) (In the ASHRAE experiment conducted in Pakistan in the summer of 1993, the PMV value reached 4.4 due to high average temperature of 40.5 °C [27]), and prediction results of the combined method are strictly controlled within the ASHRAE scale, so compared with PMV, the results predicted by the combined method can always be reasonably explained; The PMV input is limited to these six parameters, but the thermal sensation is determined by various aspects, the combined method can use multiple parameters to

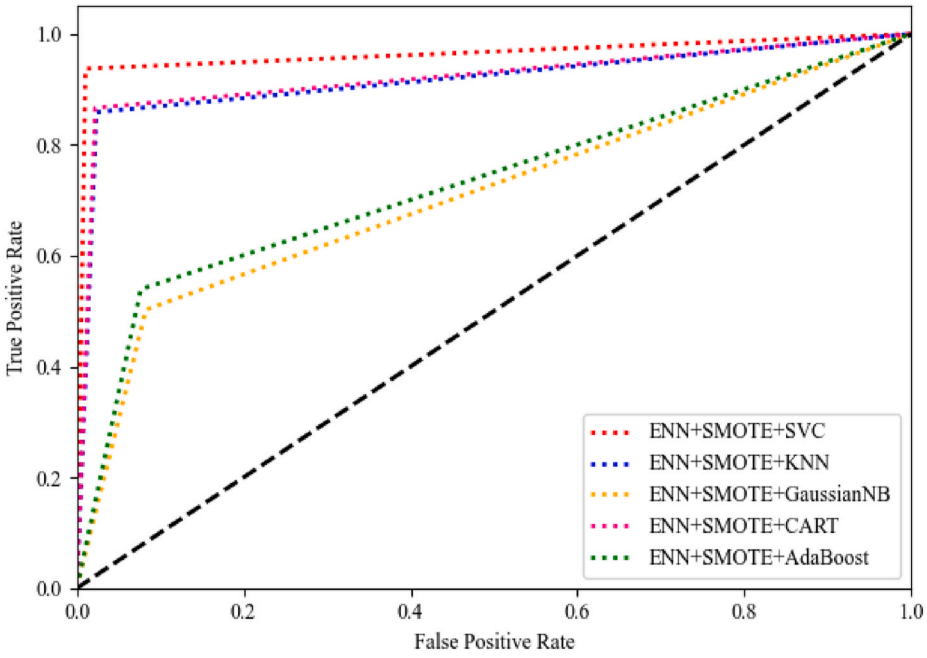


Fig. 11. ROC curves under different classification algorithms.

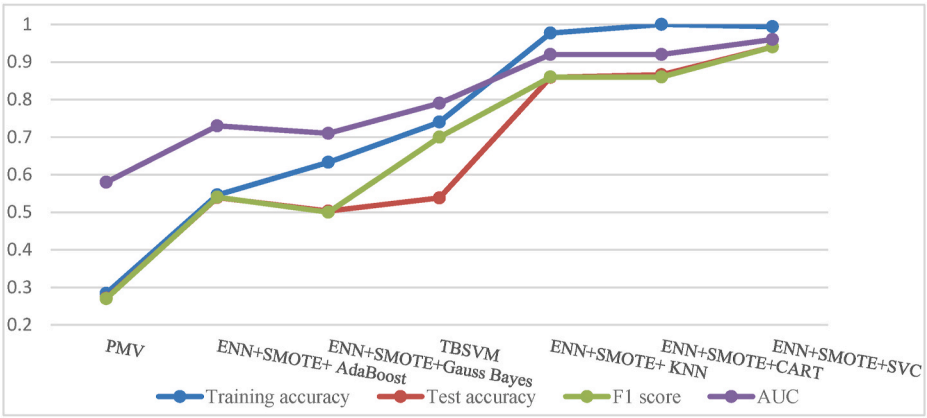


Fig. 12. Effects of different algorithms and predictor on model performance.

get more reasonable results. Therefore, the ENN + SMOTE + SVC method is more suitable for predicting daily thermal sensation than PMV. To verify this view, a comparative experiment based on the PMV and combined ENN + SMOTE + SVC method to predict thermal sensation was proposed. According to the literature [31], if the difference between the PMV value and the occupant's TSV is less than or equal to 0.5, predictions made using PMV are considered accurate. Fig. 10 depicts the ROC curve of thermal sensation prediction generated by the combined ENN + SMOTE + SVC method and PMV index. It can be clearly seen from Fig. 10 that the combined ENN + SMOTE + SVC method has obvious advantages. By comparing the evaluation indicators of PMV and SVC in Table 6, the training accuracy is 0.284, the test accuracy is 0.271, the F1 score is 0.27 and the AUC is 0.58 of PMV, respectively, which can be concluded that PMV algorithm is not suitable for the study of field observation.

Different classification algorithms have different advantages and disadvantages, as well as different applications. To verify that the SVC algorithm used in this paper is superior than other classic classification algorithms for thermal sensation prediction models, a comparative experiment based on k-Nearest Neighbour (KNN), Classification and Regression Tree (CART), AdaBoost and SVC algorithms is proposed. For the validity of the results, the data is processed the same: the total data needs to be cleaned with ENN algorithm, and the training set needs SMOTE method to achieve sample equalization. In addition, in order to verify the performance of the proposed ENN + SMOTE + SVC method and improved SVM algorithm, a comparative experiment based on TBSVM (one against one) is conducted. The comparison results are shown in Table 6, and Fig. 12 is a visualize tendency of the results. Fig. 11 shows the ROC curve of each modelling method. As can be seen from Fig. 11, the SVC algorithm is superior to other algorithms, followed by KNN and CART

algorithms, both of which have similar generalization effects. Although the training accuracy of CART algorithm is slightly higher than that of SVC algorithm, the training accuracy rate often cannot represent the generalization ability of the model, and other indicators need to be measured: SVC algorithm has a test accuracy with +0.074, F1 score with +0.08, AUC with +0.04 than CART algorithm. Among many algorithms for predicting thermal sensation models, the SVC algorithm is superior to other classification algorithms.

6. Conclusion

This paper studies the potential of the combined ENN + SMOTE + SVC method to predict thermal sensation vote. The data analysed came from the RP-884 Adaptive Model Project. Datasets with thermal sensation of a standard 7-value are selected from the project, covering more than 7000 relevant data points. The dataset is randomly divided into two parts: the training set (80%) and the test set (20%).

The main results of the analysis can be summarized as follows. The correlation matrix is used to select input features, compared to directly selecting conventional 6 parameters of PMV index as input variables, the performance of the model after training is slightly improved, indicating the effectiveness of the feature selection method. According to the comparison of the model performance with different number of features, the top 11 variables in the correlation are selected as the input of the prediction model. When the parameter pair $C = 1, \gamma = 0.1$, the test accuracy is 0.579, F1 score is 0.58, and AUC is 0.75; since the training accuracy and test accuracy of the model are low, the ENN algorithm is used to clean the data, and the generalization ability of the model has been greatly improved: the test accuracy is 0.895, the F1 score is 0.90, and the AUC is 0.94; The problem of sample imbalance in the data may also affect the performance of the model. Using the SMOTE algorithm on the training set to solve this problem, and the performance of the model is improved accordingly: the test accuracy is increased by 0.027, the F1 score is increased by 0.02, and AUC increased by 0.02; Finally, the SVC parameter C and kernel parameters γ in a are adjusted by "Grid Search", and the optimal pair($C = 4$, $\gamma = 0.0625$) is selected to train the model, and the model performance is optimized: the teat accuracy is 94.0%, the FI score is 0.94, and the AUC is 0.96.

In order to verify the superiority of the ENN + SMOTE + SVC method proposed in this paper, two sets of comparative experiments were proposed, namely ENN + SMOTE + SVC and PMV index, SVC and classic classification algorithms (KNN, Bayes, CART, AdaBoost, TBSVM). Comparative experiments found that the PMV method is not suitable for the study of field observations; the SVC algorithm is more superior than other classification algorithms in modelling a thermal sensation prediction.

In order to improve the generalization ability of the model, this paper introduces data cleaning and sample equalization processing methods. According to the size of the sample number, different forms of ENN method can be selected for data cleaning. At present, there are few researches on the problem of data itself in thermal sensation research. Therefore, this modelling method proposes a new way to improve the performance of thermal sensation prediction by introduction of data preprocessing.

Author statement

Tingzhang Liu: Supervision, Funding acquisition, Conceptualization, Methodology, Review & Editing.
Linyi Jin: Methodology, Software, Formal analysis, Data Curation, Writing - Review & Editing.
Chujun Zhong: Methodology, Writing - Original Draft, Validation.
Fan Xue: Visualization, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N.E. Klepeis, W.C. Nelson, et al., The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants, J. Expo. Anal. Environ. Epidemiol. 11 (3) (2001) 231–252.
[2] M. Frontczak, P. Wargocki, Literature survey on how different factors influence human comfort in indoor environments, Build. Environ. 46 (2011) 922–937.
[3] K. Seyboth, L. Beurskens, O. Langniss, R.E. Sims, Recognising the potential for renewable energy heating and cooling, Energy Pol. 36 (7) (2008) 2460–2463.
[4] C. Dai, H. Zhang, E. Arens, Z. Lian, Machine learning approaches to predict thermal demands using skin temperatures: steady-state conditions, Build. Environ. 114 (2017) 1–10.
[5] L. Yang, H. Yan, J.C. Lam, Thermal comfort and building energy consumption implications - a review, Appl. Energy 115 (2014) 164–173.
[6] P.O. Fanger, Analysis and Applications in Environmental Engineering, Danish Technical Press, 1970.
[7] J. von Grabe, Potential of artificial neural networks to predict thermal sensation votes, Appl. Energy 161 (2016) 412–424.
[8] S. Jing, B. Li, R. Yao, Exploring the "black box" of thermal adaptation using information entropy, Build. Environ. 146 (2018) 166–176.
[9] J.F. Nicol, M.A. Humphreys, Thermal comfort as part OF a self-regulating system, Build Res Pract 1 (1973) 173–178.
[10] K.J. McCartney, J. Fergus Nicol, Developing an adaptive control algorithm for Europe, Energy Build. 34 (6) (2002) 623–635.
[11] S. Roaf, F. Nicol, M. Humphreys, P. Tuohy, A. Boerstra, Twentieth century standards for thermal comfort: promoting high energy buildings, Architect. Sci. Rev. 53 (1) (2010) 65–77.
[12] J.T. Kim, J.H. Lim, S.H. Cho, G.Y. Yun, Development of the adaptive PMV model for improving prediction performances, Energy Build. 98 (2015) 100–105.
[13] P.O. Fanger, Bioengineering, Thermal Physiology and Comfort, 1981, https://doi.org/10.1016/S0166-1116(08)71091-4.
[14] G.S. Brager, G. Paliaga, R. De Dear, B. Olesen, J. Wen, F. Nicol, M. Humphreys, Operable windows, personal control, and occupant comfort, Build. Eng. 110 (2) (2004) 17.
[15] W.W. Liu, Z.W. Lian, B. Zhao, A neural network evaluation model for individual thermal comfort, Energy Build. 10 (39) (2007) 1115–1122, 2007.
[16] W. Li, J. Zhang, T. Zhao, J. Wang, R. Liang, Experimental Study of Human Thermal Sensation Estimation Model in Built Environment Based on the Takagi–Sugeno Fuzzy model[J], Building Simulation12, 2019, pp. 365–377.

[17] B. Salehi, A.H. Ghanbaran, M. Maerefat, Intelligent models to predict the indoor thermal sensation and thermal demand in steady state based on occupants' skin temperature, Build. Environ. 2 (169) (2020) 106579, 2020.
[18] A.C. Megri, I.E. Naqa, F. Haghighat, A learning machine approach for predicting thermal comfort indices, Int. J. Vent. 3 (2005) 363–376, 2005.
[19] L. Jiang, R. Yao, Modelling personal thermal sensations using C-Support Vector Classification (C-SVC) algorithm, Build. Environ. 99 (2016) 98–106.
[20] K.R. Jayadeva, S. Chandra, Twin support vector machines for patter classification, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 905–910.
[21] Z. Wang, Y.H. Shao, L. Bai, C.N. Li, L.M. Liu, N.Y. Deng, Insensitive stochastic gradient twin support vector machines for large scale problems, Inf. Sci. 462 (2018) 114–131.
[22] Q. Fan, Z. Wang, D. Li, D. Gao, H. Zha, Entropy-based fuzzy support vector machine for imbalanced datasets, Knowl-Based Syst 115 (2017) 87–99.
[23] Parashjyoti Borah, Deepak Gupta, Robust twin bounded support vector machines for outliers and imbalanced data, Appl. Intell. 51 (2021) 5314–5343, 2021.
[24] Z. Zhou, Machine Learning, 2010, https://doi.org/10.1093/bioinformatics/btq112 arXiv:0-387-31073-8.
[25] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Transactions on Systems, Man and Cybernetics SMC- 2 (3) (1972) 408–421, https://doi.org/10.1109/TSMC.1972.4309137.
[26] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[27] R. De Dear, Macquarie university's ASHRAE RP-884 adaptive model projectdata downloader. https://www.sydney.edu.au/architecture/staff/homepage/richard_de_dear/ashrae_rp-884.shtml, 2012.
[28] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems), third ed., 2011.
[29] F. Frauke, Igel Christian, Evolutionary tuning of multiple SVM parameters, Neurocomputing 64 (2004) 107–117.
[30] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, Acm Transactions on Intelligent Systems and Technology 2 (3) (2011) 27, https://doi.org/10.1145/1961189.1961199.
[31] R. Rana, B. Kusy, R. Jurdak, J. Wall, W. Hu, Feasibility analysis of using humidex as an indoor thermal comfort predictor, Energy Build. 64 (2013) 17–25.

Nomenclature

Abbreviation

ASHRAE: American Society of Heating Refrigerating and Airconditioning Engineer SVM: Support Vector Machine
SVC: Support Vector Classifier
ENN: Edited Nearest Neighbour
SMOTE: Synthetic Minority Oversampling Technique
KNN: K Nearest Neighbours
RBF: Radial Basis Function
HVAC: Heating, Ventilation and Air Conditioning

Symbols

PMV: Predicted Mean Vote
PPD: Predicted Percentage of Dissatisfied
ET: Effective Temperature
SET: Standard Effective Temperature
TSV: Thermal Sensation Vote
PRXY_TSA: Thermal acceptability
MET: Average metabolic rate of subject
CLO: Ensemble clothing insulation
UPHOLST: Insulation of the subject's chair
TAAV: Average of three heights' air temperature
TRAV: Average of three heights' mean radiant temperature
VELAV: Average of three heights' air speed
RH: Relative humidity
day06_ta: Outdoor 6 a.m. (min) air temp on day of survey
day15_ta: Outdoor 3 p.m. (max) air temp on day of survey
day06_rh: Outdoor 6 a.m. (max) relative humid on day of survey
day15_rh: Outdoor 3 a.m. (min) relative humid on day of survey
acc: accuracy
F1 score: harmonic average of precision and recall
ROC: Receiver Operating Characteristic
AUC: Area under Curve
R^2: Coefficient of determination
MSE: Mean Squared Error
C: the penalty coefficient of SVC
γ: RBF kernel function parameter